DOCUMENT RESUME

ED 428 083 TM 029 483

AUTHOR Brooks, Gordon P.
TITLE Precision Efficacy Analysis for Regression.
PUB DATE 1998-10-00
NOTE 65p.; Paper presented at the Annual Meeting of the Mid-Western Educational Research Association (Chicago, IL, October 1998).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Monte Carlo Methods; *Prediction; *Regression (Statistics); *Sample Size; Simulation
IDENTIFIERS Cross Validation

ABSTRACT

When multiple linear regression is used to develop a prediction model, sample size must be large enough to ensure stable coefficients. If the derivation sample size is inadequate, the model may not predict well for future subjects. The precision efficacy analysis for regression (PEAR) method uses a cross- validity approach to select sample sizes such that models will predict as well as possible in future samples. Previous studies have shown the sample sizes suggested by the PEAR method to be superior to other methods in limited cross-validity shrinkage to acceptable a priori levels. A Monte Carlo study was conducted to verify the PEAR method further for the selection of regression sample sizes and to extend the analysis to include an investigation of the effects of multicollinearity on coefficient estimates obtained through multiple linear regression analysis. Appendixes show the derivation of the PEAR method for sample size selection, and give correlation matrices, stem-and-leaf plots, and histograms of cross-validity for the study. (Contains 10 tables, 4 figures, and 116 references.) (SLD)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

Running Head: PEAR

ED 428 083

# Precision Efficacy Analysis for Regression

Gordon P. Brooks

Ohio University

TM029483

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Gordon Brooks

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Paper presented at the meeting of the Mid-Western Educational Research Association, Chicago, IL (1998, October)

Abstract

*When multiple linear regression is used to develop a prediction model, sample size must be large enough to ensure stable coefficients. If the derivation sample size is inadequate, the model may not predict well for future subjects. The precision efficacy analysis for regression (PEAR) method uses a cross-validity approach to select sample sizes such that models will predict as well as possible in future samples.*

*Previous studies have shown the sample sizes suggested by the PEAR method to be superior to other methods in limiting cross-validity shrinkage to acceptable a priori levels. The purpose of this paper is (a) to verify further the PEAR method for the selection of regression sample sizes and (b) to extend the analysis to include an investigation of the effects of multicollinearity on coefficient estimates obtained through multiple linear regression analysis.*

## Precision Efficacy Analysis for Regression

For both statistical and practical reasons, researchers should choose for their sample size "the smallest number of cases that has a decent chance of revealing a significant relationship if, indeed, one is there" (Tabachnick & Fidell, 1989, p. 129). When generalizability is the primary concern, this concept translates as the smallest sample that will provide the reliability of results required across multiple samples. Especially in multiple linear regression, which is used for many purposes, necessary sample size depends heavily on the goals and design of the analysis. "At one extreme, the null hypothesis $\rho = 0$ can often be tested powerfully with only a few dozen cases. At the other extreme, hundreds or thousands of cases might be needed to accurately estimate the sizes of higher-order collinear interactions" (Darlington, 1990, p. 380).

Several methods currently exist to help researchers narrow the choice of sample size a little more than either dozens or thousands, including conventional rules, statistical power methods, and cross-validation methods. Unfortunately, because of difficulties and contradictions among these various methods, sample size selection in multiple linear regression has been problematic (Wampold & Freund, 1987). For example, how does one reconcile the difference between Cohen's (1988) statistical power method that recommends 48 subjects, Park and Dudycha's (1974) method that advises 93 subjects, and Stevens' (1996) 15:1 subject-to-predictor ratio that suggests 60? See Table 1 for several such discrepancies. Consequently, the selection of adequate and appropriate sample sizes is not always an easy matter in multiple linear regression.

Table 1
Sample Sizes at Two Levels of Expected Sample ($R^2$) and Four Predictors

| | Assumed Population Squared Correlation | |
|---|---|---|
| Method | $R_E^2 = .25$ | $R_E^2 = .10$ [a] |
| Cohen (1988) [$1-\beta = .90, \alpha = .05$] | 48 | 144 |
| Darlington (1990) Precision Analysis [b] | 166 | 230 |
| Gatsonis & Sampson (1989) [$1-\beta = .90, \alpha = .05$] | 55 | 165 |
| Milton (1986) [$t = 2, \Delta r^2 = .02, \alpha = .05$] | 155 | 185 |
| Park & Dudycha (1974) [$\gamma_2 = .90$] [c] | 93 | 173 |
| PEAR method [$\epsilon = .22R_E^2$] | 141 | 414 |
| Predictive Power Method [$\epsilon = .20R_E^2$] | 124 | 364 |
| 15:1 (Stevens, 1996) | 60 | 60 |
| 30:1 (Pedhazur & Schmelkin, 1991) | 120 | 120 |
| $50 + 8p$ (Green, 1991) | 82 | 82 |
| Sawyer (1982) [$K = 1.05$] | 55 | 55 |

[a] for Cohen and Gatsonis & Sampson, actually $R_E = .30$. [b] for $R_E^2 = .25$, lower confidence limit (LCL) is .16; for $R_E^2 = .10$, $LCL = .04$. [c] for $R_E^2 = .25$; $\epsilon = .05$; for $R_E^2 = .10$, $\epsilon = .03$.

## Statement of the Problem

For whatever reasons, empirical study into power, generalizability, and sample size for multiple linear regression has been lacking. Subject-to-predictor conventions have existed for decades with little empirical or mathematical support. Previous work has found sample size conventions overly simplistic and very limited in their value (Brooks & Barcikowski, 1994, 1995; Drasgow & Dorans, 1982). Additionally, sample size methods offered by Park and Dudycha (1974), Cohen (1988), Gatsonis and Sampson (1989), and Sawyer (1982) were each found inadequate by Brooks and Barcikowski (1994, 1995) in some way, especially in regard to generalizability.

The general purpose of this study is to verify further a method by which the relative generalizability of sample multiple linear regression results may be analyzed. This method for assessing generalizability, called Precision Efficacy Analysis for Regression (PEAR), serves as the foundation for a method of determining appropriate sample sizes in multiple linear regression (i.e., the PEAR method). The evolution of the PEAR method extends from earlier work done by Brooks and Barcikowski (1994, 1995, 1996).

The PEAR method uses a cross-validity approach to the selection of multiple linear regression sample sizes so that regression models will predict as well as possible for future subjects. The method,

which is based on an algebraic manipulation of a cross-validation shrinkage formula, enables researchers to limit the expected shrinkage of $R^2$. Essentially, the method uses an effect size to determine the subject-to-variable ratio appropriate for the squared multiple correlation expected in a given study. For example, using one set of criteria at an expected $\rho^2$ of .40, the PEAR method suggests a subject-to-variable ratio of approximately 15:1; but at an expected $\rho^2$ of .20, the PEAR method recommends a ratio of 38:1 (see Table 2). Table 2 also shows that the PEAR method simplifies to the same subject-to-variable ratio for all numbers of variables; whereas a different ratio is required when only the number of predictors is considered (as is the case with subject-to-predictor ratios).

Table 2
Sample Sizes from the PEAR Method for Several Effect Sizes and Several Predictor Set Sizes

| | | Expected Sample Squared Multiple Correlation ($R_E^2$) | | | | | |
|---|---|---|---|---|---|---|---|
| | Predictors | .10 | .20 | .30 | .40 | .50 | .60 |
| Subjects per Predictor | | | | | | | |
| | 2 | 124.23 | 56.05 | 33.32 | 21.95 | 15.14 | 10.59 |
| | 6 | 96.62 | 43.59 | 25.91 | 17.08 | 11.77 | 8.24 |
| | 10 | 91.10 | 41.10 | 24.43 | 16.10 | 11.10 | 7.77 |
| | 14 | 88.73 | 40.03 | 23.80 | 15.68 | 10.81 | 7.56 |
| | 18 | 87.42 | 39.44 | 23.45 | 15.45 | 10.65 | 7.45 |
| Subjects per Variable [a] | | | | | | | |
| | ALL | 82.82 | 37.36 | 22.21 | 14.64 | 10.09 | 7.06 |

Note. The PEAR method is explained in detail later (here, $\epsilon = .22R_E^2$).
[a] number of variables is p+1, where p is the number of predictors.

Previous studies by Brooks and Barcikowski (1994, 1995) have compared the sample sizes suggested by the PEAR method to statistical power methods (Cohen, 1988; Gatsonis & Sampson, 1989), conventions (Green, 1991; Pedhazur & Schmelkin, 1991; Stevens, 1996), and cross-validity methods (Park & Dudycha, 1974; Sawyer, 1982). The PEAR method has been found to be superior to these existing methods in reliably limiting cross-validity shrinkage to specific acceptable a priori levels. The first problem to be studied here will be the efficiency of the PEAR method at several levels of accuracy. Investigation of this problem will help to validate the PEAR method for more extensive use with standard multiple linear regression. Further, examination of this problem may help to provide some indication as to whether certain criteria used in the PEAR method are better able to recommend adequate sample sizes

than others.

Second, the study will investigate whether the larger samples recommended by the PEAR method provide more reliable regression coefficients even under multicollinearity conditions. Although multicollinearity is known to impact the results of multiple linear regression analyses, very little is known about the effect an adequate sample size will have on multicollinear data. That is, multicollinearity has been shown consistently to be a problem when sample sizes are small, especially relative to the number of predictors. Indeed, one solution to the problem of multicollinearity is to collect additional data. Investigation of this problem will help to determine whether the use of adequate sample sizes chosen at the beginning of a study, as determined by the PEAR method, will alleviate much of the variance inflation problem associated with multicollinearity in multiple linear regression studies.

## Delimitations and Limitations of the Study

The study must be viewed from certain perspectives, which imply specific delimitations and limitations for the study. This study applies to standard regression analysis, where all predictors are entered simultaneously. More specifically, the current research proceeds based on the general linear model and multiple linear regression based upon the ordinary least squares criterion, used for prediction from a random model perspective.

Multiple linear regression is used primarily for two purposes, explanation and prediction, which as general categories include many other functions (e.g., see Afifi & Clark, 1990; Chatterjee & Price, 1991; Hocking, 1976; Montgomery & Peck, 1992; Myers, 1990). Regression can be used to explain by (a) identifying regressor variables that best explain, through their individual relative effects, the amount of a dependent variable, or (b) building models that clarify or describe the nature of the relationships among the variables. Or regression can be used to predict a score on the dependent variable for a given individual with as little error as possible. Practical application is the main emphasis of regression analysis used in prediction studies. A researcher desires to develop an efficient regression equation that optimally combines predictor scores in order to predict accurately a subject's score on a particular criterion variable (Afifi & Clark, 1990). The choice of predictors is determined primarily by their potential effectiveness in enhancing the prediction of the dependent variable. The most common, and

among the most important, use of regression equations in the social and behavioral sciences is probably prediction (Huberty, 1989; Weisberg, 1985).

The study also assumes that data follow a joint multivariate normal distribution from a random model approach. There are two models that can be used in regression analyses (Brogden, 1972; Sampson, 1974). The distinction between the two regression models is essentially between planned (fixed) and observed (random) regressor scores (Darlington, 1990). The fixed model assumes that the researcher is able to select or control the values of the independent variables before measuring subjects on the random dependent variable. From a random model perspective, both the predictors and the criterion are sampled together from what is usually assumed to be a joint multivariate normal distribution. When the predictors are random variables, they can change from one study to another (Snyder & Lawson, 1993). Because the unplanned possible scores lead to more variation than if the predictor scores are fixed, the standard errors of the regression coefficients are higher when scores are random, which causes such results as cross-validity estimates that are expected to be lower (Darlington, 1990).

The random model is usually more appropriate for social scientists because they typically measure random subjects on predictors and a criterion simultaneously and therefore are not able to fix the values for the independent variables (Berry, 1993; Brogden, 1972; Cattin, 1980b; Claudy, 1972; Darlington, 1990; Drasgow, Dorans, & Tucker, 1979; Herzberg, 1969; Park & Dudycha, 1974; Stevens, 1986, 1996). For more complete discussion of the two models, the reader is referred to Afifi and Clark (1990), Brogden (1972), Brooks (1998), Claudy (1978), Dunn and Clark (1974), Johnson and Leone (1977), and Sampson (1974).

## Fundamentals of Precision Efficacy Analysis for Regression

The primary goal of precision efficacy analysis is to reduce the upward bias of $R^2$, thereby better estimating both $\rho^2$ and $\rho_c^2$ so that results are less likely to be sample specific. The PEAR method provides researchers with a means to determine the optimum minimum sample size for prediction studies. Provided that the researcher can make a reasonable estimate of the population $\rho^2$, the PEAR method has been shown to provide very consistent precision efficacy rates.

Precision Efficacy

The term precision efficacy (*PE*) is proposed to indicate how well a regression model is expected to perform when applied to future subjects relative to its effectiveness in the derivation sample. It should be noted that Brooks and Barcikowski (1994, 1995, 1996) have used the terms "predictive power" and "precision power" for this expectation. However, it is believed that the use of the word "power" may mislead researchers into thinking that precision power is directly related to statistical power. Therefore, for the present study, the term precision efficacy will be used, recognizing that efficacy is the "the power to produce an effect" (Woolf, 1975, p.362).

Precision efficacy provides a measure of the relative efficiency of a regression equation, but does not indicate the value of a model in any absolute sense for either prediction or explanation. The formal definition of precision efficacy is

$$PE = \frac{R_C^2}{R^2} \tag{1}$$

where $R^2$ is the sample coefficient of determination and $R_C^2$ is the sample cross-validity estimate. For example, if 48% cross-validity shrinkage from sample $R^2 = .50$ to $R_C^2 = .26$ occurs, the precision efficacy is $PE = .26/.50 = .52$. Larger precision efficacy values imply that a regression model is expected to generalize better in future samples.

Cross-validity estimates describe how well a multiple linear regression equation will generalize to other samples. Several authors have described the difference between the sample $R^2$ and the cross-validity estimate $R_C^2$ as a loss in predictive power (e.g., Cattin, 1980a; Stevens, 1996). Although useful in some contexts, the absolute loss in predictive power, $(R^2 - R_C^2)$, does not provide any sense of the magnitude of loss as compared to the original sample $R^2$ value. For example, a loss in predictive power of .20 suggests drastically different results and implications for generalizability if $R^2 = .50$ $(R_C^2 = .30)$ than if $R^2 = .25$ $(R_C^2 = .05)$. Because they desire a regression model that predicts well in subsequent samples, researchers hope to limit shrinkage as much as possible relative to the sample $R^2$ value they attained.

The relationship of precision efficacy to sample size selection can be inferred and adapted from

an example used by Stevens (1996, p. 100). With a larger sample, precision efficacy would be larger because less shrinkage occurs with larger samples, all else remaining constant. Using Stevens' example, a 62% shrinkage from $R^2 = .50$ to $R_C^2 = .191$ occurs with a sample size of 50; when the sample is increased to 150, there is only a 16% shrinkage from $R^2 = .50$ to $R_C^2 = .421$. The precision efficacy in the first case would be $.191/.50 = .382$ and precision efficacy in the second case is .842.

Proportional Shrinkage. The precision efficacy formula can be manipulated algebraically into the formula $PE = 1 - (R^2 - R_C^2)/R^2$. The fraction in this equation, or proportional shrinkage, is the amount of shrinkage that occurs in $R^2$ after a cross-validity estimate, $R_C^2$, is calculated from the data relative to the $R^2$. Proportional shrinkage (*PS*) is therefore calculated by:

$$PS = \frac{R^2 - R_C^2}{R^2}. \qquad (2)$$

The precision efficacy of the regression equation, and therefore an estimate of the model's generalizability, also can be computed as $PE = 1 - PS$. For example, if sample $R^2 = .50$ and $R_C^2 = .26$, the precision efficacy for that regression model can also be described as $PE = 1 - (.50 - .26)/.50 = .52$. Proportional shrinkage of .48, and therefore precision efficacy of .52, suggests rather limited generalizability for the regression model because the $R^2$ value shrank by almost half. Lower proportional shrinkage and higher precision efficacy values imply that a regression equation is expected to generalize better in future samples relative to the model's ability to predict in the derivation sample.

Effect Size

Stevens (1996), based on analysis of Park and Dudycha's (1974) tables, has emphasized that the magnitude of the population squared multiple correlation, $\rho^2$, "strongly affects how many subjects will be needed for a reliable regression equation" (Stevens, 1996, p. 125). Similarly, Huberty (1994) noted that based on analysis of shrinkage results that "it is perhaps clear that the magnitude of $R^2$ should be considered in addition to *N/p* ratios when assessing the percent of shrinkage of $R^2$ that would result in the estimation process. That is, a general rule of thumb for a desirable *N/p* ratio (say, 10/1) may not be applicable across many areas of study" (p. 356). Indeed, all methods that account for effect size agree:

as effect size decreases, sample size must increase proportionately (e.g., Cohen, 1988; Darlington, 1990; Milton, 1986; Park & Dudycha, 1974; Gatsonis & Sampson, 1989).

Effect size enables a researcher to decide a priori not only what size relationship will be necessary for statistical significance, but also what relationship should be considered for practical significance (Hinkle & Oliver, 1983; Light, Singer, & Willett, 1990). Therefore, the first task in any sample size analysis generally is regarded to be the identification of the expected magnitude of the multiple correlation in the population. However, as Schafer (1993) wrote: "if one knew the answer to that question one would not need to do the study, but a value is needed anyway" (p. 387). Light, Singer, and Willett (1990) offered as a starting point that this effect size should be "the minimum effect size you consider worthy of your time" (p. 194). For example, because under 10% explained variance may not provide any new knowledge in the field, a researcher may choose a minimum practical effect size may be 20%. In multiple linear regression, however, the researcher must remember the effects of shrinkage. That is, if a researcher chooses 20% explained variance (i.e., $R^2 = .20$) as a minimum practical effect worthy of study, that researcher does not want a corrected sample estimate (e.g., $R_A^2$ or $R_C^2$) to be .05.

There are three basic strategies for choosing an appropriate effect size: (a) use effect sizes found in previous studies, (b) decide on some minimum effect that will be practically significant, or (c) use conventional small, medium, and large effects. No matter how it is chosen, effect size must be chosen a priori. In many cases, the researcher may have some basis for deciding the smallest correlation that would be interesting to find (practical significance), based perhaps on prior research or experience (Schafer, 1993; Shaver, 1993).

Although it is not recommended generally (e.g., Kirk, 1996; Shaver, 1993), researchers who find it difficult to hypothesize a specific effect size often rely on conventional values recommended by applied statisticians. For example, Cohen (1988) has defined conventional effect sizes for fixed model multiple linear regression such that a small effect is $R^2 = .02$, a medium effect is $R^2 = .13$, and a large effect is $R^2 = .26$. Thompson (1993) noted that empirical meta-analytic research has led to conclusions similar to Cohen's (1988) regarding typical effect sizes in multiple linear regression research. Schafer (1993) suggested that any effect less that $\rho^2 = .10$ may be too small to be of other than

theoretical interest. Stevens (1986) has suggested that $\rho^2 = .50$ is a reasonable guess for social science research; Rozeboom (1981), however, wrote that he believed $\rho^2 = .50$ to be an upper limit. Indeed, because an effect of $\rho^2 = .25$ seems unreasonably large to Schafer (1993), he recommended that it serve as an upper limit only as a last resort, when no other rationale is available. Light, Singer, and Willett (1990) echoed Schafer: "meta-analyses often reveal a sobering fact: effect sizes are not nearly as large as we all might hope" (p. 195).

Shrinkage Tolerance

Darlington (1990) defined validity shrinkage as the difference between a regression's apparent validity, for example $R^2$, and its actual predictive validity in the population, which is estimated by $R_C^2$. Stevens (1996) called this a "loss in predictive power" and it was called a "loss in $R^2$ for prediction" by Montgomery and Peck (1992). Simply put, validity shrinkage is the size of the decrease in the sample $R^2$ when an appropriate cross-validity formula is applied. The development of the PEAR method for calculating sample sizes uses this concept of validity shrinkage as a measure of a priori acceptable shrinkage tolerance, $\epsilon$. Thus, shrinkage tolerance can be defined mathematically as

$$\epsilon = R^2 - R_C^2 \tag{3}$$

which is the numerator of the proportional shrinkage fraction described in Equation 2. Shrinkage tolerance can be considered either absolute or relative. In an absolute sense, $\epsilon$ can be set to a specific value regardless of the effect size expected in a given study. That is, no matter what $R^2$ is to be used, the researcher may wish that the expected shrinkage be within .10 of the sample $R^2$ value. For example, if $R^2$ is expected to be near .50 and the researcher has chosen $\epsilon = .10$, $R_C^2$ is expected to be near .40; but if $R^2$ is expected to be near .35, the researcher is willing to accept .25 for the expected shrunken $R_C^2$ value when $\epsilon$ is set to .10.

The formula for calculating precision efficacy can also be written as $PE = 1 - \epsilon / R^2$. For example, setting the predetermined acceptable shrinkage level at $\epsilon = .20$ provides precision efficacy of .80. To carry the example out fully, precision efficacy of .80 indicates that the sample was large enough to allow the sample $R^2$ to shrink by only 20%. To provide a numerical example, if the population $\rho^2$ is thought to be .50 and $\epsilon$ is set at $.2R^2$, the sample $R^2$ is expected to shrink only by 20% to $R_C^2 = .40$

and hence precision efficacy of .80; whereas, if expected $R^2$ is near .35, $R_C^2$ would be expected near .28—again $PE = .80$. Or if $\epsilon$ is set at $.3R^2$, a sample $R^2$ of .50 will be expected to shrink 30% to $R_C^2 = .35$, a $PE$ of .70.

Solving $PE = 1 - \epsilon / R^2$ for $\epsilon$, and replacing $R^2$ with an expected, a priori $R_E^2$, results in the formula:

$$\epsilon = R_E^2 - (PE \times R_E^2) \quad (4)$$

where $R_E^2$ is an expected sample $R^2$ effect size value, chosen by the researcher perhaps based on previous research. Using this formula, a specific level of precision efficacy can be set a priori to determine the acceptable shrinkage tolerance to use in selecting an adequate sample size. For example, if the researcher wishes to obtain a cross-validity estimate expected to be not less than 80% of the sample $R^2$, a priori precision efficacy would be .80. If the expected sample $R^2$, $R_E^2$, is thought to be .50, then the shrinkage tolerance can be found by substituting the appropriate values into Equation 4. That is, shrinkage tolerance $\epsilon$ would be found a priori for this example by calculating

$\epsilon = .50 - (.80 \times .50) = .50 - .40 = .10$.

Brooks and Barcikowski (1997) determined that a slight modification to Equation 4 may provide better results when an estimated population $\rho^2$ is used with the PEAR method. The PEAR method was derived based on the use of an expected $R^2$ value rather than an estimated population $\rho^2$ value. Consequently, slightly larger than desired sample sizes are recommended when an estimated $\rho^2$ is used in the PEAR method formula and in Equation 4 (as was the case in Brooks & Barcikowski, 1994, 1995). That is, because the sample $R^2$ usually is a positively biased estimate of $\rho^2$, when the lower estimated $\rho^2$ is used in Equation 4, the $\epsilon$ value obtained is usually smaller than what would be obtained with the larger expected $R^2$. Because the PEAR method requires division by $\epsilon$, a smaller $\epsilon$ results in a larger sample size recommendation.

Hoping to compensate for this effect when $\rho^2$ is used, Brooks and Barcikowski (1997) found that a slight increase in the shrinkage tolerance $\epsilon$ did indeed provide better results for the full model, standard regression case. This adjusted $\epsilon$ is calculated by

$$\epsilon = \rho_E^2 - (PE - .1PS)\rho_E^2 \quad (5)$$

where $PS = 1 - PE$ and $\rho_E^2$ is the estimated population $\rho^2$ value (e.g., an $R_A^2$ found by the researcher in previous research or through meta-analysis). Using the same example from above results in the following: $\epsilon = .50 - ([.80 - .1(.20)] \times .50) = .50 - [.78 \times (.50)] = .11$.

In another example, Brooks (1998) showed that when $R_E^2 = .25$, using Equation 4 for $\epsilon$ resulted in a recommendation of 155 subjects. However, when $\rho_E^2 = .25$ in Equation 5, a sample size of 141 subjects was suggested. Note that when $\rho_E^2 = .25$, however, the expected $R^2$ is really $R_E^2 = .269$ (based on a formula in Herzberg, 1969); if .269 is used for $R^2$ in Equation 4, it results in the same (rounded) sample size of 141. The use of Equation 5 will be important to the current Monte Carlo study because, for the data to be generated, the population $\rho^2$ will be known but an expected $R^2$ value will not be available. A more detailed explanation can be found in Brooks (1998).

PEAR Method. Brooks and Barcikowski (1995) developed a sample size formula they called the precision power method, but within the current study will be called the PEAR method. The PEAR Method sample size formula was developed based on a cross-validity formula by Lord (as cited in Uhl & Eisenberg, 1970): $R_C^2 = 1 - (N+p+1)(1-R^2)/(N-p-1)$, where $N$ is sample size, $p$ is the number of predictors, and $R^2$ is the actual sample value. Uhl and Eisenberg (1970, p. 489) found this "relatively unknown formula" (their interpretation of Lord, 1950, differs from others) to give accurate estimates of cross-sample shrinkage, regardless of sample size and number of predictors. Algebraic manipulation of the Lord formula to solve for sample size yields the Precision Efficacy Analysis for Regression sample size formula for multiple linear regression (see Appendix A for the algebraic derivation):

$$N = (p+1) \times \frac{(2 - 2R_E^2 + \epsilon)}{\epsilon} \quad (6)$$

where $p$ is the number of predictors, $R_E^2$ is the expected sample $R^2$, and $\epsilon$ is an acceptable a priori amount of expected shrinkage. The $R_E^2$ serves as an effect size and $\epsilon$ allows researchers to decide how closely to estimate $\rho_C^2$, either as an absolute amount of acceptable shrinkage (e.g., $\epsilon = .05$) or a proportional decrease (e.g., $\epsilon = .2R_E^2$, which represents validity shrinkage of 20% from $R_E^2$ to $R_C^2 = .8R_E^2$).

When using an estimated $\rho^2$, however, $\rho_E^2$ should be used in place of $R_E^2$ in Formula 6 and Equation 5 should be used to calculate the shrinkage tolerance value $\epsilon$ (see Brooks, 1998). The resulting formula is:

$$N = (p+1)\times\frac{(2-2\rho_E^2+\epsilon)}{\epsilon} \tag{7}$$

For example, using Equation 5 provides $\epsilon = \rho_E^2 - [.80 - .1(.2)]\rho_E^2 = .22\rho_E^2$ at $PE = .80$. Based on this shrinkage tolerance level, $\epsilon = .22\rho_E^2$, the PEAR method (Formula 7) for $PE = .80$ when $\rho_E^2$ is used simplifies to:

$$N \geq \frac{2-1.78\rho_E^2}{.22\rho_E^2}\times(p+1) \tag{8}$$

The theory underlying the PEAR method for sample size selection is that the researcher, knowing that the application of an appropriate cross-validity formula is likely to cause shrinkage in $R^2$, can set a limit as to the amount of shrinkage expected to occur. Similarly, Stevens (1996), while analyzing Park and Dudycha's tables, used the example that if .40 is substituted for $R^2$ in the Stein cross-validity formula, it can be determined that "more than 15 subjects per predictor will be needed to keep the shrinkage fairly small" (p. 125), while fewer than that will be needed in $R^2 = .70$. The effect size, $\rho_E^2$ or $R_E^2$, and the shrinkage tolerance, $\epsilon$, serve as means by which the researcher can manipulate the formula in order to, in Stevens' terms, "keep the shrinkage fairly small."

Examples of the PEAR Method. By making adjustments in the shrinkage tolerance, $\epsilon$, the PEAR method may be simplified in several ways. The shrinkage tolerance, which in function is similar to the error tolerance level used in the Park and Dudycha (1974) method, must be calculated for the given specifications and the appropriate expected $R^2$ value must be determined. For example, if a researcher wanted an $R_C^2$ estimate to be at least 87% of the expected sample $R_E^2$ of .53 with four predictors, the researcher would set $PE$ to .87 and calculate $\epsilon$ from Equation 4 to be $\epsilon = .53 - (.87\times.53) = .069$. These values would then be substituted into the PEAR method formula (Equation 6) to calculate the necessary sample size as $N = 5\times[2-2(.53)+.069]/.069 = 73.12$. Therefore, at least 74 subjects should provide a large enough sample so that $R_C^2$ is expected to be greater than .46, which is 87% of the

assumed $\rho^2$ of .53. More examples of the method can be found in Brooks and Barcikowski (1996) and Brooks (1998).

## Review of Relevant Literature

Because the Precision Efficacy Analysis for Regression (PEAR) method for choosing sample sizes was developed primarily from a cross-validity perspective, the literature review will contain a review of shrinkage and cross-validation literature. Another section will address problems associated with the various existing methods of selecting sample sizes for multiple linear regression. Finally, the evolution of the PEAR method for sample size selection in multiple regression will be traced briefly.

This study will investigate, among other things, the impact of multicollinearity on multiple linear regression results. In particular, the question as to whether a proper sample size set a priori can help minimize the effects of multicollinearity. Therefore, a review of the relevant literature in the area of multicollinearity will be made, with special emphasis on issues related to sample size and the methodology that will be employed in the study.

### Generalizability and Statistical Significance

Unfortunately, many researchers apparently hold erroneous beliefs that smaller calculated probability values mean that "increasingly greater confidence can be vested in a conclusion that sample results are replicable" (Thompson, 1996, p. 27; see also Carver, 1993; Kirk, 1996; Shaver, 1993; Snyder & Lawson, 1993). Statistical significance indicates neither the magnitude nor the importance of a result (Shaver, 1993). Indeed, with a large enough sample size, a significant result may be obtained even though there is very little relationship between the criterion and the predictor variables (Asher, 1993; Snyder & Lawson, 1993).

In particular, multiple linear regression can result in a model being statistically significant, but which model provides unrealistic estimates for the relationships under investigation. The process of maximizing the correlation between the observed and predicted criterion scores requires mathematical capitalization on chance sampling error variation. When the regression equation is used with a second sample from the same population, it is most likely that the model will not perform as well as it did in the original sample; consequently, the estimate of the population multiple correlation will decrease in the

second sample (Barcikowski, 1980). For example, Stevens (1996, p. 120) provided an example regression analysis that resulted in statistical significance for the $R^2$ value of .61 (p = .036). However, when the sample $R^2$ is corrected for bias with an adjusted $R^2$ formula by the statistical computer program, the $R^2$ was decreased to $R_A^2 = .46$. Further, if a cross-validity estimate is applied to those results, the $R_C^2$ value is only .16! Clearly, the sample size ($N = 15$) used in the analysis was not adequate to produce generalizable results, but did produce statistical significance.

Sample sizes for multiple linear regression, particularly when used to develop prediction models, must be chosen so as to provide adequate power both for statistical significance and also for generalizability of the model (Barcikowski, 1980). In particular, when multiple linear regression is used to develop a prediction model, sample size must be large enough to ensure stable coefficients that will generalize from one sample to another. It is well-documented and unfortunate that many researchers do not heed this guideline. Possibly more tragic are the cases where researchers have used a groundless convention to choose their sample sizes, have ignored effect size completely, or have neglected to report an appropriate shrunken $R^2$; these studies probably provide inaccurate conclusions regarding the topics under investigation.

From a statistical power perspective, a study with an insufficient sample size stands a large chance of committing a Type II error. From a generalizability viewpoint, an insufficient sample leads to results that may apply only to the current sample and will not be useful or practical for application to other samples; that is, the correlation statistics obtained are guaranteed to be a maximum only for the particular sample from which it was calculated. In either case, time, effort, and money would have been spent arriving at results "that are inconclusive at best and which may delay further investigation of a potentially fruitful field at worst" (Streiner, 1990, p. 618).

While Darlington's (1990) simple rule that more is better certainly is true for the sake of generalizability, for the sake of practicality, there should be a caveat regarding the cost of obtaining the "more." For example, Olejnik (1984) suggested that researchers "use as many subjects as you can get *and* you can afford" [italics added] (p. 40). Streiner (1990) suggested that it is equally wasteful to study more subjects than are needed as it is not to study enough. Light, Singer, and Willett (1990) added that

"you need to know not just that 'more is better'; you need to know 'how many is enough'" (p. 186). The ability of the PEAR method to set an a priori precision efficacy level assists researchers with both concerns, from a perspective of generalizability.

Shrinkage

The importance of sample size in regression is not immediately obvious—after all, researchers have shrinkage and cross-validity formulas available to correct for inadequate sample sizes. However, a prediction model produced using a larger sample size will estimate better both the population squared multiple correlation, $\rho^2$, (using $R_A^2$) and the population squared cross-validity coefficient, $\rho_C^2$, (using $R_C^2$). For example, the true $\rho_C^2$ value for the Stevens (1996) example cited above is probably larger than .16; indeed, the true $\rho^2$ may be larger than .46—the small sample size limited the accuracy of these estimates.

Because $R^2$ is a positively biased estimator of both $\rho^2$ and $\rho_C^2$, such that $E(R^2) > \rho^2 > \rho_C^2$, researchers must report an appropriate shrunken $R^2$ (that is, $R_A^2$ or $R_C^2$) for their intended purposes (Cattin, 1980b; Claudy, 1978; Darlington, 1990; Herzberg, 1969; Hocking, 1976; Huberty & Mourad, 1980; Montgomery & Peck, 1992; Thompson, 1993). For example, "although we may determine from a sample $R^2$ that the population $R^2$ is not likely to be zero, it is nevertheless not true that the sample $R^2$ is a good estimate of the population $R^2$" (Cohen & Cohen, 1983, p. 105). The population coefficient of determination, $\rho^2$, is the unknowable squared multiple correlation that would be obtained between the criterion variable and the regression function if both are measured in the population (Herzberg, 1969; Stevens, 1996). Because this parameter is useful in describing the strength of the relationship between a criterion and a set of regressors, it is of particular interest in explanatory research (Kromrey & Hines, 1995). The most common formula used to correct $R^2$ to estimate the squared population multiple correlation is attributed most frequently to Wherry (e.g., Norusis & SPSS Inc., 1993; Dixon, 1990; Ray, 1982). The Wherry formula for adjusted $R^2$, denoted $R_A^2$, is $R_A^2 = 1-(N-1)(1-R^2)/(N-p-1)$. For example, a researcher who calculates a sample $R^2 = .3322$ with 121 subjects and 3 predictors might use an adjusted $R^2$ formula to conclude that, in the population, the multiple correlation between the criterion and the predictors is approximately $\rho = .56$, since $R_A^2 = .3151$.

Most questions concerning explanation, description, and causal analysis require an estimate of $\rho^2$, while most questions of prediction concern $\rho_c^2$. But as Herzberg (1969) noted, "in applications, the population regression function can never be known and one is more interested in how effective the *sample* regression function is in *other* samples" (p. 4). Mosteller and Tukey (1968) wrote:

> Users have often been disappointed by procedures, such as multiple regression equations, that "forecast" quite well for the data on which they were built. When tried on fresh data, the predictive power of these procedures fell dismally. . . . No one knows how to appraise a procedure safely except by using different bodies of data from those that determined it. In other words, appraisal requires some form of cross-validation. (p. 110)

Or as Cohen (1990) stated, "the investigator is not interested in making predictions for that sample—he or she *knows* the criterion values for those cases. The idea is to combine the predictors for maximal prediction for *future* samples" (p. 1306). Therefore, researchers must use and report strategies that actually do evaluate the replicability of their results. Replication is essential to confidence in the reliability or reproducibility of a result, as well as to conclusions about generalizability (Asher, 1993; Shaver, 1993). The best way to gauge this generalizability is through an estimate of $\rho_c^2$.

Cross-validity correction formulas, which are based on estimates of the mean squared error of prediction (Darlington, 1968, 1990; Herzberg, 1969), provide more accurate estimates than does $R^2$ of the squared population cross-validity coefficient, $\rho_c^2$. The cross-validity coefficient indicates how well a regression model may predict in subsequent samples because it is considered to be the multiple correlation between the actual population criterion values and the scores predicted by the sample regression equation when applied either to the population or to another sample (Cattin, 1980b; Huberty & Mourad, 1980; Kennedy, 1988; Schmitt, Coyle, & Rauschenberger, 1977).

Formula methods of cross-validity are often preferred to empirical cross-validation (e.g., data-splitting) so that the entire sample may be used for model-building. Indeed, several common formula estimates have been shown superior, or at least equivalent, to empirical cross-validation techniques (Cattin, 1980a, 1980b; Drasgow, Dorans, & Tucker, 1978; Kennedy, 1988; Morris, 1981; Rozeboom, 1978; Schmitt, Coyle, & Rauschenberger, 1977). Many such cross-validity formulas have been proposed

(Browne, 1975; Darlington, 1968; Herzberg, 1969; Lord, 1950; Nicholson, 1960; Rozeboom, 1978; Stein, 1960). When shrinkage is calculated through the use of a cross-validity formula, any finite sample size will result in a cross-validity estimate, $R_C^2$, that is smaller than the sample squared multiple correlation, $R^2$. Similar conceptually to Cronbach's reliability coefficient alpha, cross-validity formulas attempt to estimate the average of all possible empirical cross-validations (Wherry, 1975).

For example, using the random model cross-validity estimate developed independently by Stein (1960) and Darlington (1968),

$$R_C^2 = 1 - \frac{N-1}{N-p-1} \times \frac{N-2}{N-p-2} \times \frac{N+1}{N} \times (1-R^2) \quad (9)$$

where $N$ is the sample size, $p$ is the number of predictors, and $R^2$ is the sample coefficient of determination, a researcher who calculates a sample $R_C^2 = .3322$ with 121 subjects and 3 predictors might calculate the sample squared cross-validity as $R_C^2 = .2916$. This cross-validity coefficient implies that the researcher would explain 29%, not 33%, of the variance of the criterion when applying the sample regression function to future samples. The cross-validity estimates result in more shrinkage because these cross-validity corrections, unlike adjusted $R^2$ estimates, must correct for the sampling error present in both the given present study and some future study (Snyder & Lawson, 1993).

As a final note, often, researchers are interested in prediction, but also desire to know approximately what the population $\rho^2$ is. In such a case, the investigator should report not only a cross-validity estimate $R_C^2$, but also an estimate of $R_A^2$ for descriptive purposes (Thompson, 1996). Researchers must remember that the different formulas (i.e., adjusted or cross-validity) estimate different parameters and therefore are not interchangeable. For example, in large normally distributed samples, the mean, median, and mode converge; but few would argue that these are equivalent measures of central tendency—they each describe a particular facet of the distribution. The Wherry adjustment provides better estimates of the population $\rho^2$ than does any cross-validity estimate (e.g., Carter, 1979); but as Stevens (1996) indicated, "use of the Wherry formula would give a misleadingly positive impression of the cross validity predictive power of the equation" (p. 99).

Problems in Selecting Sample Sizes in Multiple Linear Regression

There are three primary types of sample size methods available for multiple linear regression: conventional rules, statistical power approaches, and cross-validation approaches. Additionally, Darlington (1990) has proposed a method based on the precision of the estimates provided by a sample. These various methods provide diverse sample size recommendations (see Table 1). The following sections describe each briefly, with emphasis on problems associated with each.

Conventions. Because cross-validity estimates are primarily functions of sample size and the number of predictors, conventions typically are based on the premise that with a large enough ratio of subjects to predictors the sample regression coefficients will be reliable and will estimate closely the true population values (Miller & Kunce, 1973; Pedhazur & Schmelkin, 1991; Tabachnick & Fidell, 1989). Conventional rules typically take the form of a subject-to-predictor ratio, usually denoted $N:p$ or $N/p$ (e.g., Halinski & Feldt, 1970; Stevens, 1986).

A well-known convention is that the sample size in a regression should equal at least 10 times the number of regressors, a ratio of subjects to predictors of 10:1 (Knapp & Campbell-Heider, 1989). Stevens (1986) recommended a 15:1 subject-to-variable ratio, which he based primarily on an analysis of Park and Dudycha's (1974) tables. Harris (1985) noted, however, that ratio conventions clearly break down for small numbers of predictors and recommended scholars investigate the utility of a difference rule, say $N - p > 50$. Knapp and Campbell-Heider (1989) recommended a combination rule of $N > 30 + 10p$. And Sawyer (1982) has developed a formula based on limiting the inflation of an alternative to mean squared error. If the inflation factor is set to a constant as Sawyer suggested, the method simplifies to a series of conventional rules. For example, if a researcher wishes for only 5% inflation, the sample size required can be approximated by $N > 10.8p + 11.8$; whereas if the researcher is willing to allow an inflation of 10%, the necessary sample size is approximately $N > 5.8p + 6.8$. Unfortunately, perhaps the most widely used sample size convention is simply to use as many subjects as you can access (Olejnik, 1984).

The most profound problem with many conventional rules advanced by regression scholars is that they lack any measure of effect size. It is generally recognized that an estimated effect size must

precede the determination of appropriate sample size. Further, Milton (1986) has indicated that determination of sample size also requires a level of precision or confidence. Finally, conventional rules are subject to change and interpretation by their users, which has resulted in the chaos of many different rules (Milton, 1986; Knapp & Campbell-Heider, 1989). For example, Stevens (1986, 1996) is explicit how he derived his recommendation of 15:1, but Tabachnick and Fidell (1989) are not so clear how they decided upon 20:1. Over time, the evolution of these rules causes their origins and rationales to become fuzzy. For example, someone who recommended a 10:1 rule may have analyzed many datasets that coincidentally all had an $R^2$ around .50.

Statistical Power Methods. Statistical power is the probability of rejecting the null hypothesis when the null hypothesis is indeed false. Several scholars have proposed regression sample size methods based on statistical power (e.g., Cohen, 1988; Cohen & Cohen, 1983; Gatsonis & Sampson, 1989; Kraemer & Thiemann, 1987; Milton, 1986; Neter, Wasserman, & Kutner, 1990). From a statistical power perspective, multiple linear regression provides several alternative statistical significance tests that can be the basis for sample size selection. Two statistical tests are most common in practice. The first such test is the test of the whole model, or the overall or omnibus test. The second common statistical test concerns the individual regression coefficients in the model. Cohen's sample size methods are among the most familiar, due to his several texts and articles on the matter.

For prediction studies, the fundamental problem with Cohen's (1988) method, or other methods based on a statistical power approach, is that it is designed for use from a fixed model, statistical power approach. And although Gatsonis and Sampson (1989) and Darlington (1990) have recommended methods from a random model approach, their methods are also based on a statistical power approach to sample size determination. Unfortunately, statistical power to reject a null hypothesis of zero multiple correlation does not inform us how well a model will predict in other samples. That is, adequate sample sizes for statistical power tell us nothing about the number of subjects needed to obtain precise estimates of stable, meaningful regression weights (Cascio, Valenzi, & Silbey, 1978; Darlington, 1990). Tests of the individual predictors may be useful in selecting predictors to include in a final model or in a regression analysis performed to analyze variance. However, these tests are not useful for those social

scientists who wish to predict scores on some criterion or simply to describe an overall relationship.

Cross-Validity Methods. The random model of regression recognizes and accounts for extra variability because, in another replication, different values for the independent variables will be obtained (Gatsonis & Sampson, 1989). That is, it is not known which specific values for the independent variables will be sampled on successive replications. Park and Dudycha (1974) noted that such a cross-validation approach is applicable to both the random and the fixed models of multiple linear regression; however, because the fixed model poses no practical problems, they emphasized the random model.

Park and Dudycha (1974) approached their calculation of sample sizes strictly with cross-validation in mind. That is, their primary concern was the estimation of $\rho_c^2$. Although Park and Dudycha's (1974) methods are recommended by Stevens (1996), there are difficulties for their practical application. Unfortunately, their tables are limited to only a few possible combinations of sample size, squared correlation, probability, and error tolerance. Fortunately, the $\rho^2$ are among the conventional values suggested by most other scholars. The error tolerance and probability levels also represent levels that may be most practical for application by researchers. Unfortunately, however, their math is complex enough that many researchers may feel unable to derive the information they would need for the cases not tabulated. Additionally, there is no clear rationale for how to determine the best choice of either $\epsilon$ or the probability to use when consulting the tables (although Stevens, 1996, implied through example that .05 and .90, respectively, are acceptable values). Finally, despite the focus on the cross-validation of regression models, Park and Dudycha's underlying theory seems to depend upon statistical power.

Darlington (1990) has provided a different approach to the determination of sample sizes, but his goal is the same: to provide estimates of population parameters that hold up under cross-validation. Darlington recommended a Fisher *z* method that can be used (a) to find both the power of tests and the precision of estimates (through confidence intervals), (b) with any value of alpha, and (c) with tests of null hypotheses other than nonassociation. It should be noted, however, that the primary purpose of Darlington's Fisher *z* method is to determine the sample size necessary for the second, validation sample; Darlington's method does not recommend a size for the initial, derivation sample.

Darlington (1990) has presented another method for the determination of multiple linear

regression sample sizes based on the ability to determine "just how accurately ${}_{T}R$ [ρ] or ${}_{T}PR$ [true partial correlation] can be estimated with a given sample size" (Darlington, 1990, p. 390). However, Darlington's precision analysis method is derived for estimates of $\rho^2$ and not $\rho_C^2$. The table provided by Darlington (p. 391), though, is structured loosely along lines analogous to precision efficacy. For example, if the researcher assumes that ρ and adjusted *R* will be .5, Darlington provides sample sizes for an acceptable lower confidence limit of .4 (80% of the sample adjusted *R* value), .3 (60%), .2 (40%), and .1 (20%).

Evolution of the Precision Efficacy Analysis for Regression Method

Because the methods described above provide contradictory sample size recommendations and (a) oversimplify the issue, (b) are too mathematically complex for many researchers to use, (c) are not based on the random model, or (d) are concerned only with statistical power and not generalizability, Brooks and Barcikowski (1994) developed a regression sample size selection method based on Rozeboom's (1978) cross-validity formula called the predictive power method. Unfortunately, although the predictive power method had higher and more accurate precision efficacy rates than the methods with which it was compared, it suffered some of the same inconsistencies across numbers of predictors and effect sizes as did the other methods. In particular, although the relative rankings of the methods remained fairly consistent across predictors, their absolute precision efficacy rates did not (see Figure 1). Also, the precision efficacy rates of all but Cohen's method and the predictive power method generally increased as the estimated $\rho^2$ effect size increased when $R_E^2$ approximated $\rho^2$ (see Figure 1).

The primary concern of the second Brooks and Barcikowski (1995) study was to determine if the PEAR method based on the Lord cross-validity formula (as cited in Uhl & Eisenberg, 1970), or any other method, provided consistently accurate precision efficacy rates as compared to a priori values. That is, did any sample size selection method for multiple linear regression successfully limit the expected validity shrinkage regardless of the number of predictors and the assumed population $\rho^2$ value?

Using an accuracy interval of $.75 \leq PE \leq .85$, Brooks and Barcikowski (1995) determined that the PEAR method was the most consistently accurate of the methods tested. That is, in all 20 conditions where $R_E^2 = \rho^2$, the PEAR method provided precision efficacy rates within the

interval $.75 \leq PE \leq .85$. The predictive power method (Brooks & Barcikowski, 1994) provided *PE* rates within the accuracy range in 13 of the 20 conditions. The accuracy of the remaining methods was low relative to these two methods: the Park and Dudycha (1974) method was accurate for 4 conditions, Sawyer (1982) for 5 cases, the $30:1$ rule (Pedhazur & Schmelkin, 1991) for 3, the $50 + 8p$ rule (Green, 1991) for 2, the $15:1$ rule (Stevens, 1986, 1996) for 5 cases, and neither the Gatsonis and Sampson (1989) method nor Cohen's method (1988) was accurate for any of the 20 conditions. Appendix C contains stem-and-leaf plots for the distributions of the average precision efficacy rates for these 20 conditions, with the accurate results underscored. Furthermore, the methods varied considerably in precision efficacy across both the number of predictors and expected $R^2$ values (as displayed in Figure 1).

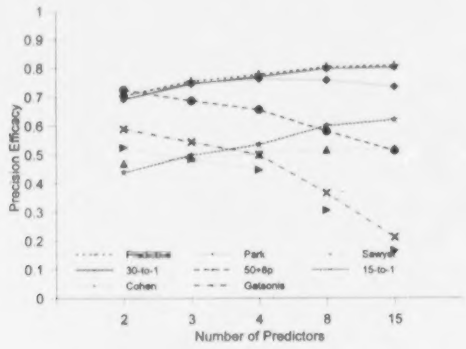
Multicollinearity

Multicollinearity, also called collinearity (e.g., Darlington, 1990; Weisberg, 1985), has been defined by Montgomery and Peck (1992) as a near linear dependence among two or more of the predictors in a regression model. More specifically, multicollinearity is the presence of substantial correlation or near linear relationship among a set of predictor variables in a regression model, such that one predictor variable may be predicted well by the other predictors (e.g., Afifi & Clark, 1984; Cohen & Cohen, 1983; Silvey, 1969). Because data are rarely orthogonal in nonexperimental research, multicollinearity is a problem of degree: multicollinearity will exist in most data to some extent (Berry, 1993; Farrar and Glauber, 1967; Montgomery & Peck, 1992; Rockwell, 1975; Willan & Watts, 1978). Indeed, Darlington (1990) indicated that partial redundancy among the predictors is the most common configuration of variables, describing it as the standard configuration. In some situations, however, the predictors may be so strongly related that the regression results are ambiguous, misleading, or erroneous (Chatterjee & Price, 1991; Montgomery & Peck, 1992).

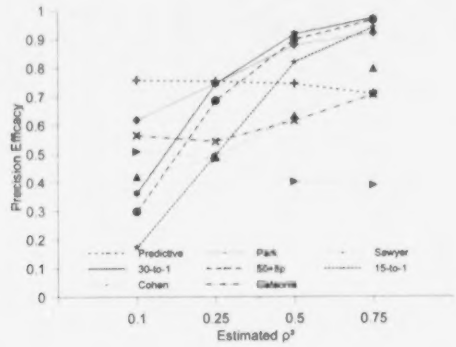
Montgomery and Peck (1992) have warned that "regression models fit to data by the method of least squares when strong multicollinearity is present are notoriously poor prediction equations" (p. 192). Darlington (1990), more optimistically, suggested that "this is an unalterable fact of life; the only solutions [to multicollinearity] lie not in cleverer analytic methods, but in such straightforward devices as

Figure 1

Precision Efficacy for nine methods at $\rho^2 = .25$ across number of predictors.




Precision Efficacy for three predictors across four effect sizes, $\rho^2$ , where estimated $\rho^2$ was equal to true population $\rho^2$.




larger sample sizes or experimental manipulation of the variables" (p. 131). Similarly, Kramer and Thiemann (1987) wrote that inclusion of several closely related predictors will decrease power and "necessitate greatly increased sample size" (p. 65). Multicollinearity is certainly a factor to be considered in multiple linear regression analyses, perhaps even in consideration of appropriate sample sizes.

The literature reveals three primary sources of multicollinearity: (a) deficient sample data, (b) model specification or overspecification, and (c) properties and characteristics of the population or the



BEST COPY AVAILABLE

process under investigation (e.g., see Berry, 1993; Chatterjee & Price, 1991; Mason, Gunst, & Webster, 1975; Montgomery & Peck, 1992. Because only singularity violates the assumptions of multiple linear regression, the ordinary least squares parameter estimates of the regression coefficients remain best, linear, unbiased estimators even in the presence of multicollinearity (Berry, 1993). However, researchers generally recognize three specific problems that result from multicollinearity: (a) interpretation of the partial coefficients because predictors duplicate each others' functions in the model, (b) sampling instability of the partial regression coefficients due to the larger size of the standard errors, and (c) model misspecification due to improper corrections to the model (Cohen & Cohen, 1983; Farrar & Glauber, 1967; Rockwell, 1975; Webster, Gunst, & Mason, 1974; Willan & Watts, 1978).

Fox (1991) has shown that, for each predictor, variance can be written as

$$Var(\beta_j) = \frac{s^2}{(N-1)s_j^2} \times \frac{1}{(1-R_j^2)} \qquad (10)$$

where $s^2$ is an estimate of *MSE* and $s_j^2$ is the estimate of the variance for the predictor coefficient, and the variance inflation factor ($VIF_j$) is $1/(1-R_j^2)$, where $R_j^2$ is the coefficient of determination obtained when predictor $j$ is regressed on the remaining $p-1$ predictors (Fox, 1991; Marquardt, 1970; Montgomery & Peck, 1992). The variance inflation factor is among the most widely recommended diagnostic techniques for detecting multicollinearity. Montgomery and Peck "believe that the *VIF*s and the procedures based on the eigenvalues of $\boldsymbol{X}'\boldsymbol{X}$ are the best currently available multicollinearity diagnostics" (p. 325).

From this equation it becomes apparent that $(1-R_j^2)$, also called tolerance, can have a significant impact on the variance of the $j$th regression coefficient—hence the name variance inflation factor. Note that Equation 10 shows that other important factors also affect the variance of the regression coefficients: sample size, estimated model error variance, and variance of the predictors themselves (Fox, 1991; Rockwell, 1975). Fox (1991) has noted that his experience suggests that "imprecise estimates in social research are more frequently the product of large error variance and relatively small samples than of serious multicollinearity" (p. 11).

## Methods

Ideally, a theoretical mathematical analysis would be offered that would describe the efficiency of the Precision Efficacy Analysis for Regression method for choosing sample sizes (Halperin, 1976; Harwell, 1990). Indeed, the efficiency of the PEAR method can be assessed analytically to some extent. Once a sample size has been chosen via the PEAR method for a given number of predictors and a given $\rho^2$, cross-validity can be estimated. For example, once the number of predictors is set at four and $\rho^2$ is assumed to be .25, the sample required by the PEAR method at $PE = .80$ is 141. Using these values in the Stein-Darlington $R_C^2$ formula gives an expected $R_C^2$ of .199, or 80% of the original $\rho^2$ value. Comparisons have been made in this way for several sample size methods in Table 3. This examination provides direct analytical evidence for the expected level of precision efficacy and therefore evidence for the adequacy of the theory underlying the PEAR method.

Table 3
Stein-Darlington Cross-Validity Estimates based on Sample Sizes from Several Methods at Two Levels of Expected Sample Squared Multiple Correlation and Four Predictors

| | $R_E^2 = .25$ | | | $R_E^2 = .10$ | | |
|---|---|---|---|---|---|---|
| Method | *N* | $R_C^2$ | *PE*[a] | *N* | $R_C^2$ | *PE*[a] |
| Cohen (1988) | 48 | .083 | .33 | 144 | .041 | .41 |
| Darlington (1990)[b] | 166 | .207 | .83 | 230 | .064 | .64 |
| Darlington (1990)[c] | 42 | .055 | .22 | 134 | .036 | .36 |
| Gatsonis & Sampson (1989) | 55 | .108 | .43 | 165 | .049 | .49 |
| Milton (1986) | 155 | .204 | .82 | 185 | .054 | .54 |
| Park & Dudycha (1974) | 93 | .171 | .68 | 173 | .051 | .51 |
| PEAR method [$\epsilon = .22\rho_E^2$] | 141 | .199 | .80 | 414 | .080 | .80 |
| Predictive Power [$\epsilon = .2\rho_E$] | 124 | .192 | .77 | 364 | .077 | .77 |
| *15:1* (Stevens, 1996) | 60 | .121 | .49 | 60 | -.054 | .00 |
| *30:1* (Pedhazur & Schmelkin, 1991) | 120 | .190 | .76 | 120 | .028 | .28 |
| $50 + 8p$ (Green, 1991) | 82 | .159 | .64 | 82 | -.009 | .00 |
| Sawyer (1982) [$K = 1.05$] | 55 | .108 | .43 | 55 | -.070 | .00 |

[a] *PE* here is calculated as $R_C^2/\rho^2$. [b] Precision Analysis. [c] Specific Conclusions.

However, several elements of the current study do not lend themselves to such analysis. For example, Mooney (1997) indicated that mathematical analysis is not possible when (a) statistical assumptions do not hold, (b) conditions required for mathematical theory are not met (e.g., the null hypothesis is known not to be true), or (c) the mathematics of the sampling distribution have not yet been worked out for a statistic. Monte Carlo methods must be used for more detailed analysis because the

sampling distribution of $R_C^2$ is very complicated and difficult to implement when $\rho^2 \neq 0$. Under the true null hypothesis, $\rho^2 = 0$, $R^2$ and $R_C^2$ obey the theoretically known distributions *F* and *t*, respectively (Herzberg, 1969). However, in the non-null case, $R^2/(1-R^2)$ has the noncentral F distribution, "which cannot be readily used for applications" (Nijsse, 1990, p. 1108; see also Fowler, 1986). Fortunately, meaningful investigation of precision efficacy rates under these conditions can be accomplished through a Monte Carlo study. As noted by Mooney (1997), Monte Carlo simulation "offers an alternative to analytical mathematics for understanding a statistic's sampling distribution and evaluating its behavior in random samples" (p. 2). That is, a Monte Carlo study can help solve problems that are mathematically intractable.

## Research Design

A Monte Carlo analysis of the precision efficacy rates of several regression sample size methods will be performed. Specifically, four methods will be compared. Three levels of precision efficacy for the PEAR method (i.e., $PE = .80$, $PE = .70$, $PE = .60$) will each be considered an individual method for the analysis. That is, given the conditions described above, Equation 7 was used to calculate sample size for the three PEAR method *PE* levels. Also for each of the conditions, the *15:1* ratio will be used to calculate sample sizes for the sake of comparison. Because a variety of factors may influence precision efficacy, three factors will be manipulated to comprise the testing situations for the present study.

First, three effect sizes that represent simultaneously the estimated population squared multiple correlation (i.e., $\rho_E^2$) and the true population $\rho^2$ will be set at: .10, .25, and .40. The numbers of predictors used to define the models in this study will be 3 predictors (i.e., 4 variables including the criterion), 7, 11, and 15 predictors. Finally, two multicollinearity conditions will be explored in the study, moderate and extensive. *Extensive* multicollinearity will be defined as over one-half of the predictors with $VIF_j > 5.0$; *moderate* multicollinearity will be defined as one-quarter of the predictors involved in such a multicollinear relationship. Two conditions where no multicollinearity exists will also be studied. Specifically, the correlation matrix for the *orthogonal* condition will contain zero correlations among the predictors. The second condition in which no multicollinearity exists will be

defined by small intercorrelations among all the predictors; numerically, for all predictors in this *non-multicollinear* condition, $VIF_j < 3.0$. Correlation matrices will be created for these conditions and treated as population correlation matrices from which multivariate normal data will be generated for each sample in the study.

A Turbo Pascal 6.0 (Borland International, Inc., 1990a) program has been written to simulate 10,000 samples for each of these 48 conditions. The program will be run as a MS-DOS application under Windows 95 on a computer equipped with an Intel Pentium-MMX 133MHz processor, which has a built-in numeric processor. Double precision floating point variables were used, providing a maximum possible range of values between $5.0 \times 10^{-324}$ to $1.7 \times 10^{308}$, stored with 15 to 16 significant digits.

During program execution, several statistics will be computed and recorded, as recommended by Harwell (1990). For each sample, the program performs a standard multiple linear regression analysis based on algorithms provided in Barcikowski (1980). The program calculates the following information from the standard, full-model regression for each sample. The statistics collected for each sample are: precision efficacy ($PE = R_C^2 / R^2$), coefficient of determination ($R^2$), Wherry adjusted $R^2$ ($R_A^2$), Stein-Darlington cross-validity $R^2$ ($R_C^2$), population $R^2$ (Herzberg, 1969), population $R_C^2$, $\rho_C^2$ (Browne, 1975), standardized regression coefficients ($\beta_j$), regression coefficients ($b_j$), standard errors of the regression coefficients ($SE_{b_j}$), and the standard error of prediction. Both $R_A^2$ and $R_C^2$ are set equal to zero when they are negative, as recommended by Cohen and Cohen (1983) and Darlington (1990). Counts are made for several statistics regarding their significance or accuracy: statistical significance for the full regression model at $\alpha = .05$, statistical significance for the regression coefficients at $\alpha = .05$, accuracy of *PE* within .05 and within 10% of a priori ($1 - PE$), accuracy of $R_A^2$ within ($.1 \times \rho^2$), and accuracy of $R_C^2$ within ($.1 \times \rho_C^2$).

In addition to these raw statistics, the appropriate calculations are made and data are collected as required for calculation of bias, *RMSE*, Relative Efficiency, statistical power, and the standard deviations of several key estimates. Statistical bias is defined as the difference between the population value $\rho^2$ and the expected value of its estimate: $Bias = E(\hat{\theta}) - \theta$, where $\theta$ is the population parameter and $E(\hat{\theta})$ is the expected value of the sample statistic or an average of the statistic over infinite samples (Drasgow,

Dorans, & Tucker, 1979; Kromrey & Hines, 1995; Mooney, 1997).

The Root Mean Squared Error (*RMSE*) provides an indication of the statistic's variability. Mean squared error is the average of the squared differences between the population parameter and its estimate for each sample. *RMSE*, then, is the square root of the mean squared error for the given statistic: $RMSE(\hat{\theta}) = \sqrt{\sum (\theta - \hat{\theta}_i)^2 / n}$, where $\theta$ is the known population parameter (as set in the computer algorithm), $\hat{\theta}_i$ is the estimate of that parameter obtained in sample $i$ of the Monte Carlo simulation, and $n$ is the total number of samples taken in the Monte Carlo study (Darlington, 1996; Drasgow, Dorans, & Tucker, 1979; Kennedy, 1988; Mooney, 1997). Mooney (1997) defined Relative Efficiency as the ratio of two *RMSE* values, multiplied by 100 to convert it to a percentage: $Relative\ Efficiency = 100 \times RMSE(\hat{\theta}_A) / RMSE(\hat{\theta}_B)$, where $\hat{\theta}_A$ and $\hat{\theta}_B$ are two different estimates the same parameter (Mooney, 1997). Values under 100 would indicate the superiority of estimator $\hat{\theta}_A$ (i.e., $\hat{\theta}_A$ with smaller *RMSE*).

Identification of the Pseudo-Population

In a Monte Carlo study, data are simulated which reflect a specified relationship among the variables (Harwell, 1990). Because this research focuses on the random model of regression, data will be generated to follow a joint multivariate normal distribution. The first step is to create population correlation matrices that meet the criteria required by this study, namely, appropriate numbers of variables, appropriate $\rho^2$ effect size values, and appropriate levels of multicollinearity. Consequently, 48 matrices will be created using these techniques.

Creation of Population Correlation Matrices. The algorithm used to create the matrices is as follows. First, for the orthogonal case, uniform random numbers between 0.0 and 1.0 are generated using a subtractive method algorithm suggested by Knuth (1981) and coded in standard Pascal by Press, Flannery, Teukolsky, and Vetterling (1989). These uniform random numbers, which are infrequently and randomly made negative, serve as possible simple correlations between the criterion and the predictors. After the first correlation is chosen, uniform random numbers are generated and chosen for the next predictors in succession based on the fact that in the orthogonal case, $R^2 = \sum (r_{yx_i})^2$ (Darlington, 1968). Once this vector of simple correlations is chosen, the remaining correlations in the matrix are set

to zero. Also, recognizing that these matrices are correlation matrices, diagonal elements are set to one.

The vector of simple correlations created in the orthogonal case is used for the remaining three multicollinearity conditions so that the simple relationships between the criterion and the predictors do not change. For the remaining cases, uniform random numbers are generated as candidate intercorrelations among the predictors. After the matrix is filled with candidates, the matrix is tested to determine whether the $\rho^2$ obtained from it meets the appropriate condition required for the Monte Carlo study, that is, within $\rho^2 \pm .005$, where $\rho^2$ is successively .40, .25, and .10.

Next, if the $\rho^2$ value falls within the required range, the matrix is then tested to determine whether it is positive definite, as is required for correlation matrices (Nash, 1990; Spath, 1992). Press, Teukolsky, Vetterling, and Flannery (1992) have suggested that the Cholesky decomposition is an efficient method for performing this test—if the decomposition fails, the matrix is not positive definite. The algorithm for the Cholesky decomposition used in this procedure was adapted from the standard Pascal code by Nash (1990). Finally, the variance inflation factors for the predictors are examined to determine if the appropriate multicollinearity condition is met. The procedure is repeated for each condition until an appropriate matrix is created for each of the 48 conditions. A Turbo Pascal 6.0 (Borland International, Inc., 1990a) program was written to generate these matrices. Appendix C contains the matrices created for three predictors across multicollinearity conditions.

Sampling Plan

After the population matrices have been created as described in the previous section, they will be used to generate sample data. More specifically, uniformly distributed pseudorandom numbers will be created to be used as input to the procedure that will convert them into multivariate normally distributed data. These procedures will be repeated as necessary for each sample created.

The L'Ecuyer (1988) generator has been chosen for present purposes. Specifically, the FORTRAN code of Press, Teukolsky, Vetterling, and Flannery (1992), has been translated into Turbo Pascal 6.0 (Borland International, Inc., 1990a) for this study. The L'Ecuyer generator was chosen because of its large period and because combined generators are recommended for use with the Box-Muller method for generating random normal deviates, as will be the case in this study (Park & Miller,

1988). The computer algorithm for the Box-Muller method to be used in this study has been adapted for Turbo Pascal 6.0, Borland International, Inc., 1990a, from the standard Pascal code provided by Press, Flannery, Teukolsky, and Vetterling, 1989.

The correlation matrices that will be created as described in a previous section will be used to generate multivariate normal data following a Cholesky decomposition procedure (also known as the square root method) recommended by several scholars (Bratley, Fox, & Schrage, 1987; Chambers, 1977; International Mathematical and Statistical Library, 1985; Karian & Dudewicz, 1991; Kennedy & Gentle, 1980; Knuth, 1981; Mooney, 1997; Morgan, 1984; Ripley, 1987; Rubinstein, 1981). Mooney (1997) has recommended that it is good practice to standardize generated variables with respect to mean and variance. Indeed, because the matrices used as input into the Cholesky procedure are correlation matrices and the means will be set to zero, the independent pseudorandom normal vectors, $\underline{X}_j$, will have means of zero and unity variances. These vectors will be generated using the implementation of the Box-Muller transformation described above.

Monte Carlo Simulations. The number of iterations for the study is based on the procedures provided by Robey and Barcikowski (1992). Significance levels for both tests on which Robey and Barcikowski's method is based were set at $\alpha = .05$ with $(1 - \beta) = .90$ as the power level; the magnitude of departure was chosen to be $\alpha \pm .2\alpha$, which falls between their intermediate and stringent criteria for accuracy. The magnitude of departure is justified by the fact that at $\pm.2\alpha$, the accuracy range for $\alpha = .05$ is $.04 \leq \alpha \leq .06$. Based on the calculations for these parameters (this set of values was not tabled), 5422 iterations would be required to "confidently detect departures from robustness in Monte Carlo results" (Robey & Barcikowski, 1992, p. 283). However, Robey and Barcikowski's method was designed to provide the number of iterations required for robustness against Type I errors; therefore, a larger number of iterations (i.e., 10,000) was chosen for the present generalizability study.

Verification of the Data Collection Procedures

According to Bratley, Fox, and Schrage (1987), verification of the algorithms should include (a) manual verification of the logic by comparing results of the computer analysis with results calculated by hand, (b) modular testing to ensure that each subroutine produces sensible output for all possible inputs,

(c) checking the results against known solutions, (d) sensitivity testing to ensure that the behavior of the computer model is sensible when parameters are varied, and (e) stress testing to ensure that strange values do not cause unexpected problems. Each of these steps was performed in preliminary analyses to verify program integrity. As changes in the program occurred as it developed, testing was repeated. Also, Type I errors were examined to test the integrity of the results from the regression algorithms used. See Brooks (1998) for a more complete description of these verification procedures.

## Data Analysis Procedures

The primary concerns of this study were (a) how appropriate are the sample sizes recommended by the PEAR method and (b) how well the PEAR method sample sizes compensate for multicollinearity. In order to answer these questions empirically, a Monte Carlo study was performed based on the design described in previous sections. The following section describes the means by which the data collected in the Monte Carlo study were analyzed.

### Problem 1: Does the PEAR method recommend appropriate sample sizes for multiple linear regression studies when cross-validity or generalizability of a prediction model is the primary purpose?

Results of the three levels of precision efficacy, that is $PE = .80$, $PE = .70$, and $PE = .60$ will be analyzed using an adaptation of the stringent accuracy criterion from Bradley (1978) and Robey and Barcikowski (1992). Specifically, bias will be calculated as the difference between actual levels of precision efficacy observed in the Monte Carlo simulation and the respective a priori $PE$ level set in the program. Based on a criterion of $PE \pm .1PS$, where $PS = 1 - PE$, bias of less than $.1PS$ will be considered accurate. For example, for a priori $PE = .70$, the result will be considered accurate if the average of observed $PE$ values over the many samples is in the range $.67 \leq PE < .73$; this criterion is equivalent to the bias criterion of $|E(PE) - PE| \leq .03$.

Examination of the bias of precision efficacy for each method (i.e., the three PEAR methods and the 15:1 ratio) will provide an estimate of how well a method performs compared to how it is expected to perform. However, in an effort to determine how the methods compare to each other, the Relative Efficiency of the methods will be compared for $PE$, $R^2$, $R_A^2$, and $R_C^2$. Comparisons of the $RMSE$ for the methods will help determine if one of the methods is preferable. Additionally, the bias of these statistics

can be compared to provide a fuller picture of the performance of the methods.

Problem 2: Does the PEAR method recommend appropriate sample sizes when multicollinearity is suspected to exist among the predictor variables included in a multiple linear regression model?

As explained in a previous section, multicollinearity is not expected to affect the values of $R^2$, $R_A^2$, or $R_C^2$. Therefore, in order to determine the effect of multicollinearity on the results obtained from a multiple linear regression analysis, the regression coefficients must be examined. The impact of multicollinearity will be examined in two ways. First, the Relative Efficiency of the methods in handling the various levels of multicollinearity will be explored. For example, the PEAR $PE = .80$ method will be compared to the PEAR $PE = .60$ method using the Relative Efficiency criterion. Specifically, the two multicollinear conditions will be compared individually with the two non-multicollinear conditions. Again, the focus will be on those predictors that are actually involved in the multicollinearity of the given predictor set.

Second, the Relative Efficiency of the regression coefficients will be isolated for each method and examined across multicollinearity conditions. For example, the comparative effect of no multicollinearity will be compared to extensive multicollinearity by analyzing the Relative Efficiency of the appropriate values for the PEAR $PE = .80$ method. It should be noted that not every regression coefficient will be involved in the multicollinearity at each level; therefore, these comparisons will focus primarily on the predictors known to be involved in multicollinear relationships.

## Results

### Problem 1

The average *PE* rates obtained for each of the PEAR method *PE* levels (i.e., .60, .70, .80) in the study are given in Table 4. Examination of Table 4 confirms that the PEAR method recommended sample sizes that provided accurate levels of precision efficacy. For all conditions tested, the PEAR method at $PE = .80$ and $PE = .70$ provided *PE* levels within the required bias criterion. For 6 of the 48 conditions, the $PE = .60$ PEAR method provided values outside of the accuracy range. Review of Table 4 also shows that the *PE* rates were more stable for higher *PE* levels (i.e., standard errors were smaller). For example, in Table 4, for $\rho^2 = .40$, $p = 11$, and the orthogonal multicollinearity

condition, the standard errors for precision efficacy at $PE = .80$ were 0.047, but for $PE = .70$ were 0.081, and for $PE = .60$ were 0.120. Table 4 also suggests that the three levels of precision efficacy each provided consistent results across numbers of predictors as well as multicollinearity conditions.

Unlike the distributions for precision efficacy, which were negatively skewed, the distributions of $R_C^2$ were relatively normal (e.g., Appendix D shows these distributions for $\rho^2 = .25$ and seven predictors in the orthogonal condition). However, the distributions clearly display the greater stability (i.e., less variability) for the $PE = .80$ level of the PEAR method as compared to the other methods.

Additional bias statistics to help distinguish the *PE* levels used with the PEAR method are provided in Table 5 for the orthogonal case. *RMSE* statistics have been provided as Table 6, also for the orthogonal multicollinearity condition. Because the correlation statistics (e.g., $R^2$, $R_A^2$, $R_C^2$) do not differ due to multicollinearity in standard full model regression, only the orthogonal cases have been tabulated.

Bias for the correlation statistics shown in Table 5 increased as the *PE* level decreased, due to the fact that smaller samples were recommended from the lower *PE* levels. For example, Table 5 shows that for $\rho^2 = .40$ and $p = 3$, the $R_C^2$ bias for $PE = .80$ was 0.029 but was 0.067 for $PE = .60$. However, because sample sizes increased, bias decreased as the effect size $\rho^2$ decreased and number of predictors increased. That is, for $p = 3$ at $\rho^2 = .40$, $R_C^2$ bias for $PE = .70$ was .050, but $R_C^2$ bias for $PE = .70$ with $p = 15$ at $\rho^2 = .40$ was .019. Similarly, Table 5 shows that bias for $p = 7$ at $\rho^2 = .40$, bias for the $PE = .70$ level was 0.029, but for $p = 7$ at $\rho^2 = .10$, bias only 0.006 for $PE = .70$. The smaller bias at lower effect sizes translates into $R_C^2$ statistics that are much closer in absolute value. For example, with $p = 7$ at $\rho^2 = .40$ in the orthogonal multicollinearity condition, $PE = .80$ resulted in average $R_C^2$ of .350 while $PE = .60$ resulted in $R_C^2 = .294$; with $p = 7$ at $\rho^2 = .10$ in the orthogonal condition, however, $PE = .80$ resulted in an average $R_C^2$ of .088, while $PE = .60$ resulted in a value of .077. This narrowing of the gap between the *PE* levels (i.e., .056 versus .011, respectively) also can be viewed graphically by examination of Figure 2, with reference to the decrease in space between the respective lines for the $PE = .80$ and the $PE = .60$ levels.

Table 4
Average Precision Efficacy (*PE*) for the Several Multicollinearity Conditions

| $\rho^2$ | Method | *p* | *N* | Orthogonal | Non | Moderate | Extensive |
|---|---|---|---|---|---|---|---|
| .40 | *PE* = .80 | 3 | 59 | .802 (.097) | .803 (.093) | .802 (.094) | .797 (.098) |
| | | 7 | 117 | .803 (.061) | .806 (.058) | .804 (.059) | .804 (.060) |
| | | 11 | 176 | .806 (.047) | .809 (.046) | .807 (.046) | .804 (.046) |
| | | 15 | 234 | .805 (.040) | .802 (.040) | .807 (.039) | .809 (.039) |
| | *PE* = .70 | 3 | 40 | .690 (.177) | .694 (.172) | .691 (.173) | .685 (.176) |
| | | 7 | 81 | .714 (.105) | .717 (.105) | .713 (.107) | .713 (.106) |
| | | 11 | 121 | .718 (.081) | .721 (.080) | .719 (.081) | .717 (.083) |
| | | 15 | 161 | .719 (.068) | .715 (.069) | .719 (.069) | .723 (.067) |
| | *PE* = .60 | 3 | 31 | .597 (.225) | .599 (.226) | .601 (.223) | .587 (.232) |
| | | 7 | 63 | .629 (.152) | .628 (.153) | .629 (.152) | .630 (.150) |
| | | 11 | 94 | .636 (.120) | .644 (.115) | .641 (.116) | .637 (.117) |
| | | 15 | 125 | .640 (.100) | .636 (.099) | .643 (.099) | .645 (.096) |
| .25 | *PE* = .80 | 3 | 113 | .800 (.087) | .805 (.083) | .799 (.089) | .808 (.081) |
| | | 7 | 226 | .802 (.054) | .798 (.055) | .802 (.054) | .803 (.054) |
| | | 11 | 339 | .805 (.042) | .803 (.042) | .801 (.043) | .808 (.042) |
| | | 15 | 452 | .803 (.036) | .803 (.037) | .802 (.037) | .808 (.036) |
| | *PE* = .70 | 3 | 77 | .698 (.155) | .702 (.153) | .697 (.155) | .707 (.151) |
| | | 7 | 153 | .708 (.100) | .702 (.101) | .710 (.096) | .710 (.096) |
| | | 11 | 230 | .715 (.074) | .716 (.072) | .710 (.076) | .721 (.072) |
| | | 15 | 307 | .718 (.062) | .717 (.062) | .715 (.063) | .723 (.060) |
| | *PE* = .60 | 3 | 59 | .605 (.209) | .609 (.207) | .602 (.210) | .616 (.203) |
| | | 7 | 117 | .621 (.142) | .615 (.145) | .624 (.140) | .624 (.139) |
| | | 11 | 176 | .634 (.108) | .632 (.106) | .627 (.109) | .640 (.104) |
| | | 15 | 234 | .637 (.090) | .635 (.092) | .633 (.092) | .643 (.088) |
| .10 | *PE* = .80 | 3 | 331 | .801 (.078) | .794 (.081) | .800 (.079) | .793 (.082) |
| | | 7 | 663 | .803 (.049) | .800 (.051) | .796 (.052) | .794 (.052) |
| | | 11 | 994 | .803 (.038) | .808 (.037) | .803 (.039) | .809 (.037) |
| | | 15 | 1325 | .803 (.033) | .800 (.033) | .801 (.034) | .809 (.032) |
| | *PE* = .70 | 3 | 222 | .697 (.145) | .687 (.152) | .694 (.147) | .686 (.151) |
| | | 7 | 444 | .711 (.089) | .706 (.090) | .700 (.094) | .696 (.095) |
| | | 11 | 667 | .714 (.068) | .721 (.065) | .714 (.068) | .725 (.063) |
| | | 15 | 889 | .715 (.058) | .711 (.059) | .712 (.059) | .723 (.056) |
| | *PE* = .60 | 3 | 168 | .600 (.200) | .588 (.209) | .598 (.201) | .585 (.205) |
| | | 7 | 335 | .622 (.131) | .617 (.132) | .608 (.138) | .603 (.139) |
| | | 11 | 503 | .628 (.102) | .638 (.097) | .633 (.098) | .643 (.095) |
| | | 15 | 671 | .634 (.083) | .629 (.084) | .630 (.085) | .646 (.079) |

Note. Standard deviations in parentheses. Average precision efficacy values that are not within the accuracy interval have been underscored to highlight them.

Although bias provides a sense of how the methods compared on average, the *RMSE* values given in Table 6 provide a better sense of how the different *PE* levels for the PEAR method performed for each sample. Specifically, the *RMSE* represents the average variation for each *PE* level for each condition. That is, whereas the bias shows the relative difference among the methods based on long run expectations (i.e., expected averages over many samples), the *RMSE* indicates how deviant on average the methods were for each sample.

For example, in Table 5 with $p = 3$ at $\rho^2 = .40$, the difference in $R_C^2$ bias between

Table 5
Bias for Orthogonal Condition

| $\rho^2$ | Method | $p$ | $N$ | $PE^a$ | $R_C^2$ | $R_A^2$ | $R^2$ |
|---|---|---|---|---|---|---|---|
| .40 | $PE$ = .80 | 3 | 59 | .002 | .029 | .006 | .005 |
| | | 7 | 117 | .003 | .019 | .005 | .005 |
| | | 11 | 176 | .006 | .013 | .002 | .002 |
| | | 15 | 234 | .005 | .011 | .002 | .002 |
| | $PE$ = .70 | 3 | 40 | -.010 | .050 | .012 | .011 |
| | | 7 | 81 | .014 | .029 | .005 | .005 |
| | | 11 | 121 | .018 | .022 | .003 | .003 |
| | | 15 | 161 | .019 | .019 | .003 | .003 |
| | $PE$ = .60 | 3 | 31 | -.003 | .067 | .016 | .015 |
| | | 7 | 63 | .029 | .041 | .007 | .007 |
| | | 11 | 94 | .036 | .033 | .005 | .004 |
| | | 15 | 125 | .040 | .029 | .004 | .003 |
| .25 | $PE$ = .80 | 3 | 113 | .000 | .018 | .003 | .003 |
| | | 7 | 226 | .002 | .011 | .002 | .002 |
| | | 11 | 339 | .005 | .007 | .002 | .000 |
| | | 15 | 452 | .003 | .007 | .001 | .001 |
| | $PE$ = .70 | 3 | 77 | -.002 | .027 | .004 | .004 |
| | | 7 | 153 | .008 | .018 | .003 | .003 |
| | | 11 | 230 | .015 | .013 | .002 | .001 |
| | | 15 | 307 | .018 | .011 | .001 | .001 |
| | $PE$ = .60 | 3 | 59 | .005 | .036 | .005 | .005 |
| | | 7 | 117 | .021 | .024 | .004 | .003 |
| | | 11 | 176 | .034 | .019 | .002 | .002 |
| | | 15 | 234 | .037 | .017 | .001 | .001 |
| .10 | $PE$ = .80 | 3 | 331 | .001 | .007 | .001 | .031 |
| | | 7 | 663 | .003 | .003 | .000 | .023 |
| | | 11 | 994 | .003 | .003 | .000 | .018 |
| | | 15 | 1325 | .003 | .002 | .000 | .016 |
| | $PE$ = .70 | 3 | 222 | -.003 | .010 | .001 | .038 |
| | | 7 | 444 | .011 | .006 | .000 | .028 |
| | | 11 | 667 | .014 | .004 | .000 | .023 |
| | | 15 | 889 | .015 | .004 | .000 | .020 |
| | $PE$ = .60 | 3 | 168 | .000 | .013 | .001 | .045 |
| | | 7 | 335 | .022 | .009 | .001 | .032 |
| | | 11 | 503 | .028 | .007 | .000 | .026 |
| | | 15 | 671 | .034 | .006 | .000 | .022 |

$PE$ = .80 and $PE$ = .60 was 0.038 (i.e., 0.067 - 0.029); but Table 6 shows that with $p$ = 3 at $\rho^2$ = .40, the difference in $R_C^2$ *RMSE* between $PE$ = .80 and $PE$ = .60 was 0.059 (i.e., .173 - .114). Similarly, the difference in $R_C^2$ *RMSE* between the two *PE* levels at $\rho^2$ = .10 with $p$ = 3 was only 0.016 (i.e., .048 - .032). The *RMSE* statistics for precision efficacy also confirm that the $PE$ = .80 level provided more stable results than the lower *PE* levels. For example, Table 6 indicates that *PE RMSE* with $p$ = 7 at $\rho^2$ = .25 was 0.054 for $PE$ = .80, but was 0.100 for $PE$ = .70 and 0.144 for $PE$ = .60.

Table 6
Average *RMSE* for Orthogonal Condition

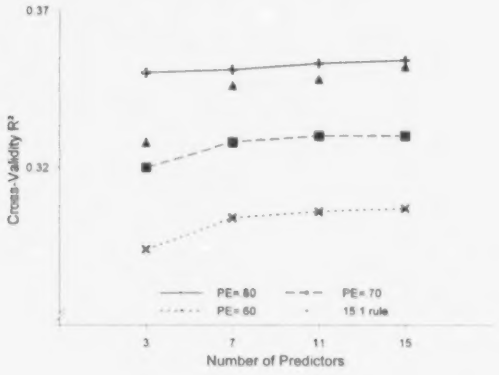
| $\rho^2$ | Method | *p* | *N* | *PE** | $R_C^2$ | $R_A^2$ | $R^2$ |
|---|---|---|---|---|---|---|---|
| .40 | $PE$ = .80 | 3 | 59 | .097 | .114 | .103 | .097 |
| | | 7 | 117 | .061 | .081 | .074 | .069 |
| | | 11 | 176 | .047 | .065 | .059 | .055 |
| | | 15 | 234 | .040 | .057 | .052 | .049 |
| | $PE$ = .70 | 3 | 40 | .177 | .149 | .128 | .118 |
| | | 7 | 81 | .106 | .103 | .090 | .082 |
| | | 11 | 121 | .083 | .084 | .073 | .067 |
| | | 15 | 161 | .071 | .073 | .064 | .058 |
| | $PE$ = .60 | 3 | 31 | .225 | .173 | .145 | .130 |
| | | 7 | 63 | .155 | .125 | .104 | .092 |
| | | 11 | 94 | .125 | .103 | .085 | .075 |
| | | 15 | 125 | .108 | .089 | .074 | .065 |
| .25 | $PE$ = .80 | 3 | 113 | .087 | .077 | .072 | .070 |
| | | 7 | 226 | .054 | .054 | .051 | .049 |
| | | 11 | 339 | .042 | .044 | .042 | .041 |
| | | 15 | 452 | .036 | .039 | .037 | .035 |
| | $PE$ = .70 | 3 | 77 | .155 | .096 | .088 | .085 |
| | | 7 | 153 | .100 | .070 | .064 | .061 |
| | | 11 | 230 | .075 | .056 | .052 | .049 |
| | | 15 | 307 | .065 | .049 | .045 | .043 |
| | $PE$ = .60 | 3 | 59 | .209 | .113 | .103 | .097 |
| | | 7 | 117 | .144 | .083 | .074 | .069 |
| | | 11 | 176 | .113 | .067 | .060 | .057 |
| | | 15 | 234 | .097 | .058 | .052 | .048 |
| .10 | $PE$ = .80 | 3 | 331 | .078 | .032 | .031 | .031 |
| | | 7 | 663 | .050 | .023 | .023 | .023 |
| | | 11 | 994 | .039 | .019 | .018 | .018 |
| | | 15 | 1325 | .033 | .016 | .016 | .016 |
| | $PE$ = .70 | 3 | 222 | .145 | .041 | .039 | .038 |
| | | 7 | 444 | .090 | .029 | .028 | .028 |
| | | 11 | 667 | .070 | .024 | .023 | .023 |
| | | 15 | 889 | .059 | .021 | .020 | .020 |
| | $PE$ = .60 | 3 | 168 | .200 | .048 | .046 | .045 |
| | | 7 | 335 | .133 | .034 | .032 | .032 |
| | | 11 | 503 | .106 | .029 | .027 | .026 |
| | | 15 | 671 | .090 | .024 | .023 | .022 |

Table 7 provides a quantitative measure, Relative Efficiency (*RE*), by which the *PE* levels can be compared for the several statistics tabulated. For example, regardless of the number of predictors, level of multicollinearity, and the $\rho^2$ value, the Relative Efficiency statistics for all three correlation statistics show that the *RMSE* of the $PE$ = .80 level was about 80% of the *RMSE* for the $PE$ = .70 level (the *RE* values were primarily in a range from about 77% to about 83%). Similarly, Relative Efficiency shows that *RMSE* of $PE$ = .70 for the correlation statistics was about 86% that of the $PE$ = .60 level. These Relative Efficiency statistics suggest that the $PE$ = .80 level of the PEAR method was about

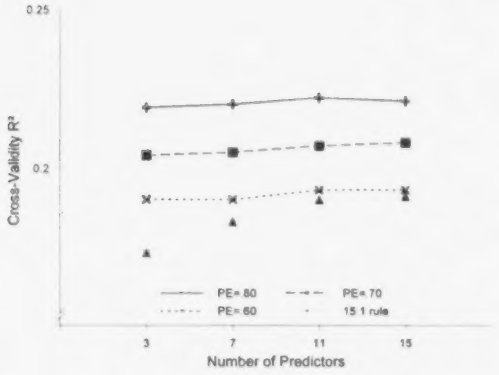
Figure 2

Average cross-validity statistics (i.e., $R_C^2$) for the three *PE* levels and the *15:1* subject-to-predictor ratio across number of predictors when:

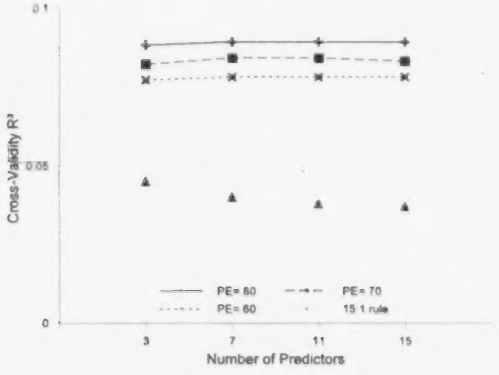
true $\rho^2$ = .40




true $\rho^2$ = .25




true $\rho^2$ = .10




BEST COPY AVAILABLE

20% more efficient than the $PE = .70$ level, which in turn was about 14% more efficient than the $PE = .60$ level. Figure 3 shows these relationships graphically by comparing each sample size method to the $PE = .80$ level of the PEAR method for one set of conditions. The Relative Efficiency of the *15:1* ratio can be seen to vary considerably depending upon the level of $\rho^2$.

Problem 2

The PEAR method has been shown to provide accurate results for the expected level of cross-validity, $R_C^2$. Indeed, results showed that not only were the estimates of $R_C^2$ stable across multicollinearity conditions, but so also were the standard errors of those estimates. For example, for $p = 7$ at $\rho^2 = .40$ and $PE = .80$, the average $R_C^2$ values were very tightly around .353 (.351 in the orthogonal condition, .355 in the non-multicollinear condition, .353 in the moderate multicollinearity condition, and .352 in the extensive multicollinearity condition); additionally, the standard deviations for those averages ranged tightly around .079 (.079, .078, .079, and .079, respectively).

However, the ability to produce a desired $R_C^2$ value does not necessarily imply that the regression weights derived for a certain model will be stable across samples. In order to determine the stability of the regression coefficients, they must be inspected individually. That is, the standard errors of the regression coefficients must be examined in order to determine the effect of varying sample sizes on the stability of the coefficients.

For the conditions with three predictors, Table 8 and Table 9 provide the standard errors of the coefficients for the four sample size methods. These tables show that the higher precision efficacy levels that recommended larger samples consistently resulted in smaller standard errors of the coefficients, regardless of the number of predictors, effect size, or multicollinearity. Although they have not been tabulated, the results showed similar patterns for the 7, 11, and 15 predictor cases as well.

Table 10 provides the Relative Efficiency of the methods compared for all numbers of predictors, all multicollinearity levels, and all effect sizes. For this table, the standard errors for the individual predictors were used for comparison because, for unbiased estimates such as the regression coefficients, *RMSE* approximates the standard error. To create Table 10, the Relative Efficiency of each predictor was calculated and then those Relative Efficiency values were averaged for the predictor set. It would

Table 7
Relative Efficiency for Orthogonal Condition

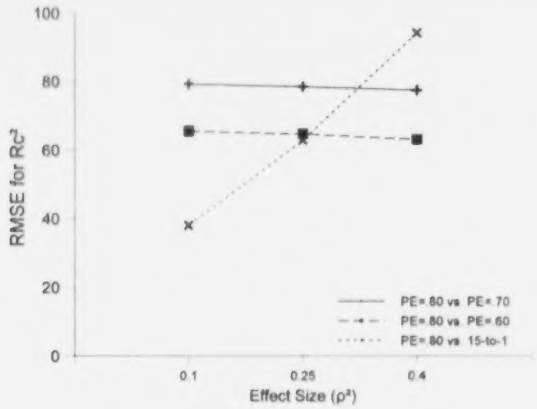
| $\rho^2$ | $p$ | Method Comparison | $PE^a$ | $R_c^2$ | $R_A^2$ | $R^2$ |
|---|---|---|---|---|---|---|
| .40 | 3 | *RMSE*(.80) / *RMSE*(.70) | 54.8 | 76.5 | 80.5 | 82.2 |
| | | *RMSE*(.80) / *RMSE*(.60) | 43.1 | 65.9 | 71.0 | 74.6 |
| | | *RMSE*(.70) / *RMSE*(.60) | 78.7 | 86.1 | 88.3 | 90.8 |
| | 7 | *RMSE*(.80) / *RMSE*(.70) | 57.5 | 79.6 | 82.2 | 84.1 |
| | | *RMSE*(.80) / *RMSE*(.60) | 39.4 | 65.6 | 71.2 | 75.0 |
| | | *RMSE*(.70) / *RMSE*(.60) | 68.4 | 82.4 | 86.5 | 89.1 |
| | 11 | *RMSE*(.80) / *RMSE*(.70) | 56.6 | 77.4 | 79.7 | 83.6 |
| | | *RMSE*(.80) / *RMSE*(.60) | 37.6 | 63.1 | 69.4 | 74.7 |
| | | *RMSE*(.70) / *RMSE*(.60) | 66.4 | 81.6 | 87.1 | 89.3 |
| | 15 | *RMSE*(.80) / *RMSE*(.70) | 56.3 | 78.1 | 81.3 | 84.5 |
| | | *RMSE*(.80) / *RMSE*(.60) | 37.0 | 64.0 | 70.3 | 75.4 |
| | | *RMSE*(.70) / *RMSE*(.60) | 65.7 | 82.0 | 86.5 | 89.2 |
| .25 | 3 | *RMSE*(.80) / *RMSE*(.70) | 56.1 | 79.4 | 81.8 | 82.4 |
| | | *RMSE*(.80) / *RMSE*(.60) | 41.6 | 68.1 | 70.6 | 72.2 |
| | | *RMSE*(.70) / *RMSE*(.60) | 74.2 | 85.8 | 86.3 | 87.6 |
| | 7 | *RMSE*(.80) / *RMSE*(.70) | 54.0 | 77.1 | 79.7 | 80.3 |
| | | *RMSE*(.80) / *RMSE*(.60) | 37.5 | 65.1 | 68.9 | 71.0 |
| | | *RMSE*(.70) / *RMSE*(.60) | 69.4 | 84.3 | 86.5 | 88.4 |
| | 11 | *RMSE*(.80) / *RMSE*(.70) | 56.0 | 78.6 | 80.8 | 83.7 |
| | | *RMSE*(.80) / *RMSE*(.60) | 37.2 | 64.7 | 70.0 | 71.9 |
| | | *RMSE*(.70) / *RMSE*(.60) | 66.4 | 82.4 | 86.7 | 86.0 |
| | 15 | *RMSE*(.80) / *RMSE*(.70) | 55.4 | 79.6 | 82.2 | 83.7 |
| | | *RMSE*(.80) / *RMSE*(.60) | 37.1 | 67.2 | 71.2 | 75.0 |
| | | *RMSE*(.70) / *RMSE*(.60) | 67.0 | 84.5 | 86.5 | 89.6 |
| .10 | 3 | *RMSE*(.80) / *RMSE*(.70) | 53.4 | 80.5 | 79.5 | 79.5 |
| | | *RMSE*(.80) / *RMSE*(.60) | 39.0 | 68.8 | 67.4 | 68.9 |
| | | *RMSE*(.70) / *RMSE*(.60) | 73.0 | 85.4 | 84.8 | 86.7 |
| | 7 | *RMSE*(.80) / *RMSE*(.70) | 55.6 | 82.8 | 82.1 | 82.1 |
| | | *RMSE*(.80) / *RMSE*(.60) | 37.6 | 70.6 | 71.9 | 71.9 |
| | | *RMSE*(.70) / *RMSE*(.60) | 67.7 | 85.3 | 87.5 | 87.5 |
| | 11 | *RMSE*(.80) / *RMSE*(.70) | 55.7 | 79.2 | 78.3 | 78.3 |
| | | *RMSE*(.80) / *RMSE*(.60) | 36.8 | 65.5 | 66.7 | 66.7 |
| | | *RMSE*(.70) / *RMSE*(.60) | 66.0 | 82.8 | 85.2 | 85.2 |
| | 15 | *RMSE*(.80) / *RMSE*(.70) | 55.9 | 76.2 | 80.0 | 80.0 |
| | | *RMSE*(.80) / *RMSE*(.60) | 36.7 | 66.7 | 69.6 | 69.6 |
| | | *RMSE*(.70) / *RMSE*(.60) | 65.6 | 87.5 | 87.0 | 87.0 |

Note. Comparisons to the *15:1* ratio were not tabulated because they are only incidental to the study.

not have been appropriate to average the results for Table 10 across predictors if the results had not been so consistent. For example, in Table 10 for $p = 3$ at $\rho^2 = .40$ in the orthogonal condition, the Relative Efficiency of the $PE = .80$ level as compared to $PE = .70$, represented as $RMSE(.80)/RMSE(.70)$, is shown to be 80.8%. Using the values from Table 8, it can be determined that for $p = 3$ at $\rho^2 = .40$ in the orthogonal condition, the Relative Efficiency for coefficients 1 was 80.9% (.102/.126); similarly, Relative Efficiency for coefficient 2 can be calculated to be 81.7% and for coefficient 3 at 79.6%.

Figure 3

*RMSE* for the PEAR method at $PE = .70$ and $PE = .60$ and the *15:1* ratio compared to the $PE = .80$ PEAR method across effect sizes, averaged for number of predictors.




There is a striking similarity between the Relative Efficiency statistics in Table 10 and those found in Table 7 for the correlation statistics. Specifically, the Relative Efficiency statistics show that, on average, the magnitude of the standard errors of the coefficients from the $PE = .80$ level were about 20% smaller than those from the $PE = .70$ level. Similarly, the comparisons of the $PE = .70$ and $PE = .60$ levels provided *RE* statistics that ranged tightly around the 86% level.

Multicollinearity is known to affect the standard errors of the regression coefficients derived for a model. Indeed, comparisons of Table 8 and Table 9 confirm that the standard errors increased not only for the predictors specifically identified as multicollinear, but also the predictors whose relationships were neither orthogonal nor multicollinear. For example, from Table 8, for $\rho^2 = .25$ for all sample size methods, the standard errors for the coefficients in the non-multicollinear condition were larger than those from the orthogonal condition for two of the three predictors. That is, the standard errors for these coefficients increased by over 70% from the orthogonal condition despite that the relationships among the predictors were known not to be multicollinear according to their variance inflation factors.

Table 8
Average Standard Errors of the Standardized Coefficients ($SE_{b_j}$) for Three Predictors for Non-Multicollinear Conditions

| | | | Orthogonal | | | Non-Multicollinear | | |
|---|---|---|---|---|---|---|---|---|
| $\rho^2$ | Method | *N* | $SE_{b_1}$ | $SE_{b_2}$ | $SE_{b_3}$ | $SE_{b_1}$ | $SE_{b_2}$ | $SE_{b_3}$ |
| .40 | *PE* = .80 | 59 | .102 | .103 | .094 | .108 | .108 | .096 |
| | *PE* = .70 | 40 | .126 | .126 | .118 | .134 | .135 | .120 |
| | *PE* = .60 | 31 | .147 | .147 | .136 | .155 | .155 | .139 |
| | 15:1 *ratio* | 45 | .119 | .118 | .109 | .127 | .126 | .111 |
| .25 | *PE* = .80 | 113 | .080 | .080 | .079 | .139 | .082 | .136 |
| | *PE* = .70 | 77 | .098 | .099 | .097 | .170 | .100 | .166 |
| | *PE* = .60 | 59 | .114 | .113 | .111 | .195 | .115 | .193 |
| | 15:1 *ratio* | 45 | .131 | .132 | .128 | .228 | .132 | .223 |
| .10 | *PE* = .80 | 331 | .052 | .052 | .050 | .071 | .066 | .055 |
| | *PE* = .70 | 222 | .064 | .064 | .062 | .089 | .083 | .068 |
| | *PE* = .60 | 168 | .074 | .073 | .071 | .101 | .095 | .079 |
| | 15:1 *ratio* | 45 | .146 | .147 | .143 | .204 | .189 | .155 |

Note. $SE_{b_j}$ approximates *RMSE* when estimate is unbiased as is $\beta_j$.

Table 9
Average Standard Errors of the Standardized Coefficients for 3 Predictors for Multicollinear Conditions

| | | | Moderate | | | Extensive | | |
|---|---|---|---|---|---|---|---|---|
| $\rho^2$ | Method | *N* | $SE_{b_1}$ | $SE_{b_2}$ | $SE_{b_3}$ | $SE_{b_1}$ | $SE_{b_2}$ | $SE_{b_3}$ |
| .40 | *PE* = .80 | 59 | .202 | .254* | .140 | .183 | .264* | .308* |
| | *PE* = .70 | 40 | .254 | .312* | .173 | .228 | .327* | .382* |
| | *PE* = .60 | 31 | .295 | .365* | .201 | .264 | .387* | .453* |
| | 15:1 *ratio* | 45 | .236 | .293* | .160 | .212 | .308* | .357* |
| .25 | *PE* = .80 | 113 | .154 | .213* | .146 | .129 | .381* | .407* |
| | *PE* = .70 | 77 | .189 | .260* | .177 | .158 | .466* | .499* |
| | *PE* = .60 | 59 | .218 | .302* | .210 | .179 | .537* | .573* |
| | 15:1 *ratio* | 45 | .252 | .349* | .239 | .209 | .631* | .672* |
| .10 | *PE* = .80 | 331 | .114 | .151* | .090 | .128* | .124* | .065 |
| | *PE* = .70 | 222 | .140 | .187* | .113 | .156* | .152* | .080 |
| | *PE* = .60 | 168 | .160 | .213* | .128 | .180* | .176* | .093 |
| | 15:1 *ratio* | 45 | .327 | .436* | .260 | .363* | .352* | .185 |

Note. $SE_{b_j}$ approximates *RMSE* when estimate is unbiased as is $\beta_j$. * indicates predictor with *VIF*> 5.0 (i.e., involved in multicollinearity).

Figure 4 shows graphically the average standard error for sets of seven predictors at $\rho^2 = .40$, $\rho^2 = .25$, and $\rho^2 = .10$. The graphs show that as multicollinearity increased, the standard errors of the coefficients increased, as an average for the sets of predictors. All of the sample size methods (i.e., three *PE* levels and the *15:1* ratio) were affected by this increase in standard error, but the effect was

Table 10
Average Relative Efficiency of the Standardized Coefficients Across Predictors

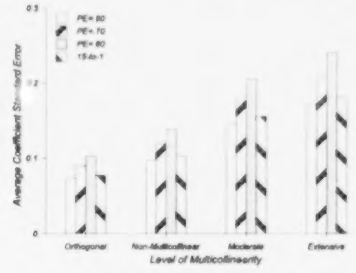
| $\rho^2$ | $p$ | Method Comparison | Orthogonal | Non | Moderate | Extensive |
|---|---|---|---|---|---|---|
| .40 | 3 | *RMSE*(.80) / *RMSE*(.70) | 80.8 | 80.2 | 80.6 | 80.5 |
| | | *RMSE*(.80) / *RMSE*(.60) | 69.5 | 69.5 | 69.2 | 68.5 |
| | | *RMSE*(.70) / *RMSE*(.60) | 86.1 | 86.6 | 85.9 | 85.1 |
| | 7 | *RMSE*(.80) / *RMSE*(.70) | 81.7 | 81.6 | 81.3 | 82.9 |
| | | *RMSE*(.80) / *RMSE*(.60) | 70.5 | 70.6 | 71.2 | 71.1 |
| | | *RMSE*(.70) / *RMSE*(.60) | 86.3 | 86.6 | 87.6 | 85.8 |
| | 11 | *RMSE*(.80) / *RMSE*(.70) | 81.4 | 81.8 | 81.6 | 80.5 |
| | | *RMSE*(.80) / *RMSE*(.60) | 70.7 | 70.4 | 70.8 | 70.5 |
| | | *RMSE*(.70) / *RMSE*(.60) | 86.8 | 86.1 | 86.7 | 87.6 |
| | 15 | *RMSE*(.80) / *RMSE*(.70) | 81.7 | 81.5 | 80.4 | 81.9 |
| | | *RMSE*(.80) / *RMSE*(.60) | 70.7 | 70.6 | 69.8 | 70.7 |
| | | *RMSE*(.70) / *RMSE*(.60) | 86.5 | 86.7 | 86.9 | 86.3 |
| .25 | 3 | *RMSE*(.80) / *RMSE*(.70) | 81.3 | 81.9 | 82.0 | 81.7 |
| | | *RMSE*(.80) / *RMSE*(.60) | 70.7 | 71.0 | 70.2 | 71.3 |
| | | *RMSE*(.70) / *RMSE*(.60) | 87.0 | 86.7 | 85.7 | 87.4 |
| | 7 | *RMSE*(.80) / *RMSE*(.70) | 81.2 | 81.5 | 81.2 | 81.6 |
| | | *RMSE*(.80) / *RMSE*(.60) | 70.0 | 71.0 | 69.9 | 70.7 |
| | | *RMSE*(.70) / *RMSE*(.60) | 86.2 | 87.1 | 86.1 | 86.6 |
| | 11 | *RMSE*(.80) / *RMSE*(.70) | 81.4 | 81.6 | 81.6 | 81.4 |
| | | *RMSE*(.80) / *RMSE*(.60) | 70.5 | 70.8 | 70.6 | 71.1 |
| | | *RMSE*(.70) / *RMSE*(.60) | 86.6 | 86.8 | 86.5 | 87.3 |
| | 15 | *RMSE*(.80) / *RMSE*(.70) | 81.8 | 81.2 | 81.0 | 81.4 |
| | | *RMSE*(.80) / *RMSE*(.60) | 71.2 | 70.6 | 70.2 | 70.5 |
| | | *RMSE*(.70) / *RMSE*(.60) | 87.0 | 86.9 | 86.8 | 86.5 |
| .10 | 3 | *RMSE*(.80) / *RMSE*(.70) | 81.0 | 80.1 | 80.6 | 81.6 |
| | | *RMSE*(.80) / *RMSE*(.60) | 70.6 | 69.8 | 70.8 | 70.5 |
| | | *RMSE*(.70) / *RMSE*(.60) | 87.2 | 87.2 | 87.9 | 86.4 |
| | 7 | *RMSE*(.80) / *RMSE*(.70) | 81.9 | 81.6 | 82.1 | 82.4 |
| | | *RMSE*(.80) / *RMSE*(.60) | 70.4 | 70.5 | 70.6 | 72.0 |
| | | *RMSE*(.70) / *RMSE*(.60) | 86.0 | 86.4 | 86.0 | 87.4 |
| | 11 | *RMSE*(.80) / *RMSE*(.70) | 81.1 | 81.9 | 81.8 | 81.7 |
| | | *RMSE*(.80) / *RMSE*(.60) | 70.4 | 70.9 | 71.2 | 70.6 |
| | | *RMSE*(.70) / *RMSE*(.60) | 86.8 | 86.6 | 87.1 | 86.5 |
| | 15 | *RMSE*(.80) / *RMSE*(.70) | 81.0 | 80.9 | 81.7 | 81.1 |
| | | *RMSE*(.80) / *RMSE*(.60) | 70.2 | 70.4 | 70.7 | 70.7 |
| | | *RMSE*(.70) / *RMSE*(.60) | 86.6 | 87.1 | 86.5 | 87.2 |

Note. $SE_{b_j}$ approximates *RMSE* when estimate is unbiased as is $\beta_j$.

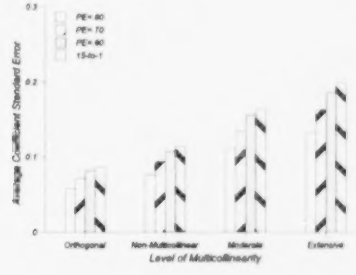
consistent when the methods were compared. Examination of Table 10 shows that the Relative Efficiency of the methods remained consistent despite the presence of multicollinearity. For example, across all effect sizes and numbers of predictors, regardless of the magnitude of standard error caused by multicollinearity, the $PE = .80$ level used with the PEAR method remained approximately 20% more efficient than the $PE = .70$ level, just as the $PE = .70$ level remained about 14% more efficient than the $PE = .60$ level.

Figure 4

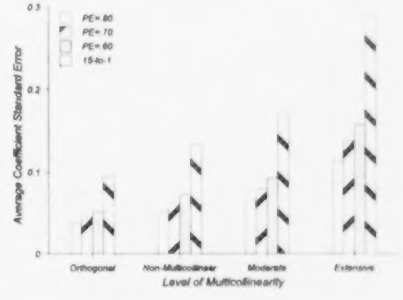
Average standard errors for the regression coefficients for three *PE* levels and the *15:1* subject-to-predictor ratio in the seven predictor conditions when true $\rho^2 = .40$.




Average standard errors for the regression coefficients for three *PE* levels and the *15:1* subject-to-predictor ratio in the seven predictor conditions when true $\rho^2 = .25$.




Average standard errors for the regression coefficients for three *PE* levels and the *15:1* subject-to-predictor ratio in the seven predictor conditions when true $\rho^2 = .10$.



Again, if these results had not been so consistent, it would not have been appropriate to group them in Figure 4. It can be seen in Table 8 and Table 9 that the standard errors for the coefficients do vary individually. However, results used to derive Table 10 confirm that, despite the differing magnitudes of the coefficient standard errors, the Relative Efficiency relationship holds true when the methods are compared. It was determined, because these Relative Efficiency held across comparisons, that averaging the standard errors for the predictor sets would not present false characterizations of the relationships among the *PE* levels as represented graphically in Figure 4.

Additionally, all sample size methods produced similar results when each was compared against itself across multicollinearity conditions. The Relative Efficiency statistics for each method compared to its orthogonal condition were similar. For example, in the moderate multicollinearity condition for Coefficient 2, which was involved in a multicollinear relationship, all sample size methods resulted in similar Relative Efficiency values near 38%. That is, for each method, the standard error from the orthogonal condition was 38% as large as the standard error for the moderate multicollinearity situation. Although Relative Efficiency was equivalent for all the *PE* levels, a review of Table 8 and Table 9 reminds the reader that standard errors were generally smaller for higher *PE* levels.

## Discussion

The results of the first research problem confirmed what Brooks and Barcikowski (1995, 1997) have found previously. That is, the PEAR method seems to provide accurate precision efficacy rates across several effect sizes and numbers of predictors. The PEAR method as defined for this study was based on an estimated population $\rho_E^2$ value rather than an expected sample $R_E^2$ value. The results suggest that the adaptation (i.e., Equation 5) of the original shrinkage tolerance formula (i.e., Equation 4) performs very well when an estimate of the population parameter is the more readily available effect size.

It should be noted, however, that from a practical perspective, the reasonable estimation of either an expected $R_E^2$ or an estimated $\rho_E^2$ is more important than which shrinkage tolerance formula is chosen. The differences from the two equations (Equation 4 and Equation 5) are minimal when compared to the differences caused by incorrect estimation of effect size. For example, Brooks and Barcikowski (1995) found that when $R_E^2 = .25$ but $\rho^2 = .10$, precision efficacy rates were in the .47 to .50 range for

$PE = .80$. Given accurate estimation of effect sizes, however, the difference in *PE* rates for Equation 4 and Equation 5 should be very small (e.g., about .02 for $PE = .80$). Consequently, whereas the more complex Equation 5 was required for the highly specific Monte Carlo simulation that used an estimated $\rho_E^2$, the use of the simpler Equation 4 often may be acceptable from a practical perspective with an expected $R_E^2$.

Bias. There seems to be a slight accuracy advantage to higher levels of precision efficacy used a priori with the PEAR method. In particular, the $PE = .80$ and the $PE = .70$ values for the PEAR method were within the acceptable bias interval for every condition, whereas the $PE = .60$ level was not accurate in six conditions. Brooks and Barcikowski's (1997) data showed that using Equation 5, the $PE = .80$ level was accurate in 94% of the cases and $PE = .70$ was accurate in 97%; the $PE = .60$ level was slightly less accurate, at 93%. When each of these a priori precision efficacy levels was not accurate in that 1997 study, the large preponderance of results were higher than expected, thereby recommending more subjects than necessary rather than less.

Indeed, although accurate for each case in the present study, most of the *PE* rates for $PE = .80$ and $PE = .70$ were in the upper half of the accuracy range (i.e., above the a priori rate); the $PE = .60$ level was also above the a priori rate for most conditions, including the six cases where it was not accurate. However, because the *PE* rates fell within the accuracy range especially at higher *PE* levels, this result provides more confidence that the expected *PE* values will be, on average, *at least as large as* the a priori *PE* level.

The benefit of an accurate method is that neither too few nor too many subjects will be recommended for a study. Although, more data is generally better, the value of obtaining additional data usually must be weighed against the opportunity cost of the extra time, effort, and expense associated with its collection. Sample size methods such as the PEAR method endeavor to recommend the minimum size for research samples. As Brewer and Sindelar (1987) wrote, "there is no such thing as a maximum sample size" (p. 75); but a recommended minimum sample size will help the researcher to determine what is necessary to achieve desired generalizability, in the case of multiple linear regression used for prediction.

Relative Efficiency. The Relative Efficiency of the different PEAR levels were investigated in hopes of detecting a best a priori level of precision efficacy. However, no clear choice was found. The $PE = .80$ level seemed to perform about 20% better than the $PE = .70$ level in the full model case, for orthogonal as well as multicollinear predictors. Similarly, the $PE = .70$ level performed about 14% better than the $PE = .60$ level across all conditions.

From these Relative Efficiency statistics it would seem that the $PE = .80$ level used with the PEAR method would be most desirable. However, more information must be considered. For example, at lower population $\rho^2$ effect sizes, the statistics based on the methods become rather close in absolute value. An example cited earlier showed that at $\rho^2 = .10$ with three predictors, $R_C^2$ was .088 for the $PE = .80$ level but only .077 for $PE = .60$. The $PE = .80$ level required 331 subjects to obtain its slightly larger $R_C^2$, whereas the $PE = .60$ level only required 168 subjects to obtain a value that many researchers might find acceptable. Other researchers may determine, however, that the additional subjects recommended by the $PE = .80$ level are well worth the added precision efficacy.

These dramatic differences in sample sizes must be balanced against the expected gain in $R_C^2$, particularly at lower effect sizes. The sample size differences are not quite so striking at higher effect sizes, but still must be considered. For example, at $\rho^2 = .40$ and three predictors, the extra 28 subjects recommended by the $PE = .80$ ($N \geq 59$) level as compared to the $PE = .60$ level ($N \geq 31$) resulted in the more noticeable difference in average $R_C^2$ of .350 versus .294, respectively. Fortunately, thoughtful adjustments to the a priori precision efficacy level or the shrinkage tolerance enable researchers to use the PEAR method to make such choices.

Recommendations. Because the $PE = .80$ and the $PE = .70$ levels of precision efficacy were slightly more accurate than the $PE = .60$ level, it is recommended that practitioners use an a priori precision efficacy value of at least $PE = .70$. In particular, the average $R_C^2$ results indicate that for moderate $\rho^2$ effect sizes (e.g., $\rho^2 = .40$ and $\rho^2 = .25$), the best choice may be $PE = .80$. That is, the $PE = .80$ level of precision efficacy keeps shrinkage to generally more acceptable absolute level. As effect size decreases, the researcher must pay closer attention to the trade-offs between relative and absolute shrinkage, as well as the opportunity costs of gathering samples of the required sizes. Finally, it

is noteworthy that although the only non-PEAR method included in the current study was the *15:1* subject-to-predictor ratio, the PEAR method once again showed its comparative value (cf. Brooks & Barcikowski, 1994, 1995).

Multicollinearity

One of the most difficult aspects in the interpretation of multiple linear regression results is the analysis of predictors that are related. In particular, situations arise in which predictors are highly correlated and multicollinearity becomes a significant problem. Many scholars have suggested that multicollinearity causes problems with the interpretation of regression results and even may affect the ability of a regression model to predict. For correlation statistics in the full model situation, multicollinearity is not an issue; that is, correlation statistics (e.g., $R^2$, $R_A^2$, $R_C^2$) are not affected by multicollinearity in the data. However, the standard errors of the regression coefficients sometimes are affected substantially by the presence of multicollinear relationships among a subset of predictors.

The results showed, however, that the Relative Efficiency of the *PE* levels chosen for the study remains surprisingly consistent across coefficients for all conditions. That is, neither the number of predictors, the effect size, nor the level of multicollinearity seemed to affect the relative performance of the standard errors of the coefficients for the different *PE* levels. Specifically, just as for the full model correlation statistics and the orthogonal coefficient standard errors, Relative Efficiency of the multicollinear coefficient standard errors was about 80% for $PE = .80$ as compared to $PE = .70$; for $PE = .70$ versus $PE = .60$, Relative Efficiency was approximately 86%. Generally speaking, the *PE* levels that recommended larger sample sizes produced smaller standard errors for the coefficients. Consequently, increasing sample size may not cure multicollinearity, but does stabilize the standard errors of the coefficients relatively.

In fact, the results that pertain specifically to the multicollinearity question suggest that when multicollinearity is suspected among a set of predictors, the $PE = .80$ level, or perhaps higher, may be the best choice. That the $PE = .80$ level of precision efficacy results in 20% more efficient coefficients than $PE = .70$ in terms of their standard errors recommends the $PE = .80$ level for use with the PEAR method. Because the standard errors become inflated, a 20% more efficient solution often might

be advantageous. For pure prediction problems, the size of the standard errors caused by multicollinearity is less worrisome; that is, even less stable coefficients (e.g., from $PE = .70$) seem to result in $R_C^2$ estimates that are just as stable as orthogonal coefficients. However, if the researcher hopes to interpret the coefficients, their standard errors will have a more significant impact.

## Conclusions

The primary goal of Precision Efficacy Analysis for Regression is to provide a means by which the researcher can assess the generalizability of a prediction model relative to its performance in the derivation sample. Precision Efficacy Analysis for Regression has been shown through several studies (Brooks & Barcikowski, 1994, 1995, 1997) to be a viable method for this generalizability analysis.

There are four primary reasons that argue for the importance of Precision Efficacy Analysis for Regression and the PEAR method of choosing sample sizes used to develop prediction models. First, precision efficacy is a means by which researchers can assess the prediction potential of a regression model relative to its performance in the derivation sample. Second, the PEAR method provides a means by which researchers can choose samples by setting a priori effect sizes, shrinkage tolerance, and precision efficacy levels. Third, results from both the present study and previous research (e.g., Brooks & Barcikowski, 1995) show that prediction models produced using appropriately large sample sizes will better estimate $\rho_C^2$. Fourth, the most important reason is that a model based on a proper sample size, as suggested by the PEAR method, will provide more reliable regression weights. Therefore, these models will predict better for future subjects because, ultimately, the efficiency of a prediction model depends not on correlation statistics such as $R_A^2$ and $R_C^2$, but on the stability of the regression coefficients.

Analysis of the results from the present study also provide evidence that the PEAR method recommends sample sizes that accurately meet the a priori expectations for precision efficacy (i.e., limit shrinkage to the levels expected). The method, which is flexible and can be adjusted based on specific research needs, provides consistent results at the three levels of a priori precision efficacy studied here. Analysis of the results has also shown that although multicollinearity tends to affect the stability of regression coefficients and regression models, the PEAR method can be adjusted in several ways to account for these differences.

The PEAR method appears to fill an important gap in the multiple linear regression literature in that it recommends sample sizes for prediction based not only on the number of predictors in a study, but also on the size of the effect expected. Indeed, most sample size methods in other areas of statistics, including fixed model regression, consider effect size to be an essential part of the calculation. Some may argue that effect sizes are too difficult to determine, but blind adherence to conventional subject-to-predictor ratios certainly cannot be better research practice. Sometimes, it is indeed difficult to determine an expected effect size—perhaps due to inadequate or unsatisfactory previous research, misinterpretation of results by other researchers, or lack of research in the topic area. When prior research is not available, pilot studies become a very important step in the research process, for pilot studies can provide an expectation of effect size. Also, careful interpretation of previous results and meta-analysis of multiple studies can help to provide at least a meaningful effect size that constitutes practical significance. When no prior knowledge is available and a pilot study is impossible, only as a last resort should conventional effect sizes be chosen.

The PEAR method can be viewed from one perspective as simply cross-validation in reverse. That is, instead of determining by how much the sample $R^2$ will shrink due to the sample size; the PEAR method determines how large a sample to use to keep $R^2$ from shrinking too much. Although at first glance the method may seem much more complex than the conventional subject-to-predictor ratios often espoused in the literature, it is not. In particular, conventional rules typically take the form $N \geq C \times p$, where $C$ is a constant based on someone's experience and $p$ is the number of predictors in the full regression model. The PEAR method, in contrast, takes the form $N \geq C \times (p + 1)$, where $C$ is variable depending on the effect size, precision efficacy, and shrinkage tolerance set by the researcher and $(p + 1)$ is the total number of variables in the model (i.e., including the criterion variable). Previous research by Brooks and Barcikowski (1995) has shown that a similar method based on $p$ (the predictive power method, Brooks & Barcikowski, 1994) does not perform as well as $(p + 1)$.

Finally, the Monte Carlo study from Brooks and Barcikowski (1994) also showed that when generalizability is the priority, one needs not worry much about statistical power. That is, when sample sizes are chosen with precision efficacy as the primary criterion, statistical power is well above the

standard .80 that is typically recommended. Indeed, for sample sizes chosen via the predictive power method, statistical power rates for cases where the $R_E^2$ approximates $\rho^2$ were over .90. Like precision efficacy, however, statistical power rates fell dramatically to unacceptable levels when $R_E^2$ overestimated $\rho^2$.

Caveats for Samples of Any Size

The use of mathematical cross-validity formulas does not supersede the need for the validation of regression models in other samples. The cross-validity formulas suggest how well a model should perform, assuming that the sample from which it was derived was reasonably representative of the population; however, any given sample can deviate from what would be expected or representative. Further, no matter what the precision efficacy, a model that does not predict well in a derivation sample also probably will not predict well in any other samples.

Developing a model with good precision efficacy should be considered only a first step in this validation process. The statistical correction cross-validity formulas attempt to predict the mean of all cross-validation attempts. Empirical cross-validation, in contrast, may result in a correlation that by chance might be lower or higher than the average of several such cross-validations (Wherry, 1975). However, the actual performance of a prediction model in a new sample (as opposed to data-splitting) provides intangible evidence not available with the use of cross-validity formulas. Further, cross-validation does not depend upon the assumptions required for use of the cross-validity equations, thus providing a possible substitute when the assumptions are not met (Darlington, 1990; Wherry, 1975). However, the PEAR method can be used to determine sample sizes even if an actual cross-validation is to be performed later; the results of such a cross-validation should be less likely to vary dramatically when based on an appropriate sample.

Also, these results are based on long-run expectations of the performance of the PEAR method. Berry (1993) noted that "unbiasedness of OLS [ordinary least squares] estimators in no way ensures that an individual estimate of a regression parameter based on a single sample will equal its population value" (p. 18). Similarly, although the expected value of precision efficacy has shown to be accurate in the long-run, any given sample size based on the PEAR method may not produce a precision efficacy value

within the stringent accuracy range used in this study. However, results based on larger samples are less likely to differ, because larger samples generally result in smaller standard errors.

Darlington (1990) and Montgomery and Peck (1992) also have expressed the importance not only of model validation (e.g., cross-validation), but also of model adequacy. According to Montgomery and Peck (1992), checking for robustness or model adequacy requires residual analyses for violations of assumptions, searching for high leverage or overly influential observations, and other analyses that test the fit of the regression model to the available data. Darlington (1990) has described robustness in this way:

> Robustness is the ability to draw valid conclusions even in the absence of standard assumptions such as normality and homoscedasticity. . . . When the assumptions of normality and homoscedasticity are not met, a study may lack robustness even when its sample size far exceeds the recommendations [for sample size]." (p. 379)

Darlington (1990) added that robustness to violations of assumptions continues to increase as sample size increases.

Further, Darlington (1990) reminded researchers that when statistical significance is found despite a small sample size, those results cannot be criticized from a statistical perspective. However, research performed in the evolution of the PEAR method has reminded researchers that such is not necessarily the case when the generalizability of results is the primary concern. That is, small samples rarely provide the generalizable prediction models that researchers might expect given the statistical significance achieved.

## Recommendations for Future Research

There are a number of issues that the present study was unable to elucidate. Therefore, the following recommendations are made for research to further investigate sample sizes for prediction models developed using multiple linear regression. First, there are aspects of multicollinearity that have not been addressed in this study. For example, results from this study were not able to describe the resulting magnitude of standard errors of the coefficients. That is, sample size alone was not enough to explain the larger or smaller standard errors of multicollinear predictor coefficients. Future research

should investigate multicollinearity as a more continuous variable. Further, future studies should examine whether larger variance inflation factors cause more dramatic inflation problems. Future research can explore questions of sample specific multicollinearity, that which changes to some degree for each sample. Also, perhaps some statistical methods for managing multicollinearity could be examined, such as stepwise regression, all-subsets regression, or ridge regression.

Second, the data in the present study were generated through computer simulation. Often, real data do not behave in the same manner as simulated data (Micceri, 1989). It may be possible to develop future studies that incorporate the use of large datasets comprised of data from real research. Having such data will allow the calculation of the true population cross-validity. Also, future studies are required to determine the efficacy of the PEAR method when the data are not distributed normally. There is reason to believe that the PEAR method, with its larger sample sizes (relative to many conventional rules) will be useful even with non-normal data. Berry (1993) noted that "as one's sample size increases, one can show decreasing concern whether the normality assumption is met" (p. 82). Other data issues include the possibility of using fixed model data (e.g., dummy variables) or the impact of heteroscedasticity on prediction may be studied.

## Epilogue

It is hoped that the method of generalizability analysis presented in this study (Precision Efficacy Analysis for Regression) and its associated sample size method will provide researchers better tools for the adequate development and design of their regression studies. The PEAR method shows much promise in providing sample sizes that keep cross-validity shrinkage to a minimum—that is, to an acceptable shrinkage tolerance level set a priori by the researcher. It is further hoped that both the evidence presented and the simplicity of the PEAR method will encourage researchers to consider more carefully the issues of sample size, effect size, and generalizability for multiple linear regression research. The results presented in this study show that the PEAR method may be useful, especially for standard full model regression, despite the presence of multicollinear predictors.

The goal of this study is to help the researcher to determine the minimum sample size required for a given prediction study, not to add another citation to the repertoire of the skeptical expert. That is,

the PEAR method can be adjusted in many ways, resulting in "just the right" number of subjects for almost any circumstance. For example, the precision efficacy level may be adjusted, or the shrinkage tolerance value may be changed, or the effect size may be altered in order to justify a given sample size after the fact.

As with any research technique, the PEAR method requires honest and a priori use to be effective: thoughtful choices are required for both effect size and precision efficacy before a sample size is calculated. Not choosing an appropriate sample size may jeopardize interpretations and conclusions from a study or may provide spurious results. "As harsh as it sounds, when researchers cannot provide an adequate sample, they should seriously consider the option of not conducting the research until an adequate amount of data is available" (Brewer & Sindelar, 1987, p. 77).

Because generalizability may be an even more important issue than statistical power in much regression research, an assessment technique such as Precision Efficacy Analysis for Regression appears beneficial to a more complete understanding of regression results. Additionally, researchers must be aware of the potential hazards of choosing an inappropriate effect size or ignoring effect size completely when selecting sample sizes. Finally, researchers must remember that no statistical analysis or adjustment (such as a cross-validity estimate) can repair problems caused by a small, nonrandom, or unrepresentative sample (Cooley & Lohnes, 1971; Miller & Kunce, 1973).

References

Afifi, A. A., & Clark, V. (1990). Computer-aided multivariate analysis (2nd ed.). New York, Van Nostrand Reinhold.

Asher, W. (1993). The role of statistics in research. Journal of Experimental Education, 61, 388-393.

Barcikowski, R. S. (1980). Regression Analysis. Unpublished manuscript, Ohio University at Athens.

Berry, W. D. (1993). Understanding regression assumptions (Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-092). Thousand Oaks, CA: Sage.

Borland International, Inc. (1990a). Turbo Pascal (Version 6.0) [Computer Software]. Scotts Valley, CA: Authors.

Bradley, J. V. (1978). Robustness? British Journal of Mathematical and Statistical Psychology, 31, 144-152.

Bratley, P., Fox, B. L., & Schrage, L. E. (1987). A guide to simulation (2nd ed.). New York: Springer-Verlag.

Brewer, J. K., & Sindelar, P. T. (1987-88). Adequate sample size: A priori and post hoc considerations. Journal of Special Education, 21, 74-84.

Brogden, H. E. (1972). Some observations on two methods in psychology. Psychological Bulletin, 77, 431-437.

Brooks, G. P. (1998). Precision efficacy analysis for regression: Development and justification of a new sample size method for multiple linear regression. Unpublished doctoral dissertation, Ohio University, Athens.

Brooks, G. P., & Barcikowski, R. S. (1994, April). A new sample size formula for regression. Paper presented at the meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. ED 412247).

Brooks, G. P., & Barcikowski, R. S. (1995, October). Precision power method for selecting sample sizes. Paper presented at the meeting of the Mid-Western Educational Research Association, Chicago, IL. (ERIC Document Reproduction Service No. ED 412246).

Brooks, G. P., & Barcikowski, R. S. (1996). Precision power and its application to the selection of regression sample sizes. Mid-Western Educational Researcher, 9(4), 10-17.

Brooks, G. P., & Barcikowski, R. S. (1997). The precision power method for the selection of regression sample sizes. Manuscript in preparation, Ohio University.

Browne, M. W. (1975). Predictive validity of a linear regression equation. British Journal of Mathematical and Statistical Psychology, 28, 79-87.

Carter, D. S. (1979). Comparison of different shrinkage formulas in estimating population multiple correlation coefficients. Educational and Psychological Measurement, 39, 261-266.

Carver, R. P. (1993). The case against statistical significance testing, revisited. Journal of Experimental Education, 61, 287-292.

Cascio, W. F., Valenzi, E. R., & Silbey, V. (1978). Validation and statistical power: Implications for applied research. Journal of Applied Psychology, 63, 589-595.

Cattin, P. (1980a). Estimation of the predictive power of a regression model. Journal of Applied Psychology, 65, 407-414.

Cattin, P. (1980b). Note on the estimation of the squared cross-validated multiple correlation of a regression model. Psychological Bulletin, 87, 63-65.

Chambers, J. M. (1977). Computational methods for data analysis. New York: John Wiley & Sons.

Chatterjee, S. & Price, B. (1991). Regression analysis by example (2nd ed.). New York: John

Wiley & Sons.

Chatterjee, S., & Yilmaz, M. (1992). A review of regression diagnostics for behavioral research. Applied Psychological Measurement, 16, 209-227.

Claudy, J. G. (1972). A comparison of five variable weighting procedures. Educational and Psychological Measurement, 32, 311-322.

Claudy, J. G. (1978). Multiple regression and validity estimation in one sample. Applied Psychological Measurement, 2, 595-607.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112, 155-159.

Cohen, J., & Cohen, P. (1983). Applied multiple regression/correlation analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cooley, W. W., & Lohnes, P. R. (1971). Multivariate data analysis. New York: John Wiley & Sons.

Darlington, R. B. (1968). Multiple regression in psychological research and practice. Psychological Bulletin, 69, 161-182.

Darlington, R. B. (1990). Regression and linear models. New York: McGraw-Hill.

Darlington, R. B. (1996). Estimating the true accuracy of regression predictions. Mid-Western Educational Researcher, 9(4), 29-31.

Dixon, W. J. (1990). BMDP statistical software manual to accompany the 1990 software release (Vol. 1). Berkeley, CA: University of California.

Drasgow, F., & Dorans, N. J. (1982). Robustness of estimators of the squared multiple correlation and squared cross-validity coefficient to violations of multivariate normality. Applied Psychological Measurement, 6, 185-200.

Drasgow, F., Dorans, N. J., & Tucker, L. R. (1979). Estimators of the squared cross-validity coefficient: A Monte Carlo investigation. Applied Psychological Measurement, 3, 387-399.

Dunn, O. J., & Clark, V. A. (1974). Applied statistics: Analysis of variance and regression. New York: John Wiley & Sons.

Farrar, D. E., & Glauber, R. R. (1967). Multicollinearity in regression analysis: The problem revisited. Review of Economic Statistics, 49, 92-107.

Fowler, R. L. (1986). Confidence intervals for the cross-validated multiple correlation in predictive regression models. Journal of Applied Psychology, 71, 318-322.

Fox, J. (1991). Regression diagnostics (Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-079). Thousand Oaks, CA: Sage.

Gatsonis, C., & Sampson, A. R. (1989). Multiple correlation: Exact power and sample size calculations. Psychological Bulletin, 106, 516-524.

Green, S. B. (1991). How many subjects does it take to do a regression analysis? Multivariate Behavioral Research, 26, 499-510.

Halinski, R. S., & Feldt, L. S. (1970). The selection of variables in multiple regression analysis. Journal of Educational Measurement, 7, 151-157.

Halperin, S. (1976, April). Design of Monte Carlo studies. Paper presented at the meeting of the American Educational Research Association, San Francisco, CA. (ERIC Document Reproduction Service No. ED 121 850)

Harris, R. J. (1985). A primer of multivariate statistics (2nd ed.). Orlando, FL: Academic Press.

Harwell, M. R. (1990, April). Summarizing Monte Carlo results in methodological research. Paper presented at the meeting of the American Educational Research Association, Boston, MA. (ERIC Document Reproduction Service No. ED 319 775)

Herzberg, P. A. (1969). The parameters of cross-validation. Psychometrika Monograph Supplement, 34 (2, Pt. 2).

Hinkle, D. E., & Oliver, J. D. (1983). How large should a sample be? A question with no simple answer? Or.... Educational and Psychological Measurement, 43, 1051-1060.

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. Biometrics, 32, 1-49.

Huberty, C. J. (1989). Problems with stepwise methods—better alternatives. In B. Thompson (Ed.), Advances in social science methodology: A research annual (Vol. 1, pp. 43-70).Greenwich, CT: JAI.

Huberty, C. J. (1994). A note on interpreting an $R^2$ value. Journal of Educational and Behavioral Statistics, 19, 351-356.

Huberty, C. J., & Mourad, S. A. (1980). Estimation in multiple correlation/prediction. Educational and Psychological Measurement, 40, 101-112.

International Mathematical and Statistical Library. (1985). Stat/PC Library. Houston, TX: Authors.

Johnson, N. L., & Leone, F. C. (1977). Statistics and experimental design in engineering and the physical sciences. New York: John Wiley & Sons.

Karian, Z. A., & Dudewicz, E. J. (1991). Modern statistical systems, and GPSS simulation: The first course. New York: Computer Science Press.

Kennedy, E. (1988). Estimation of the squared cross-validity coefficient in the context of best subset regression. Applied Psychological Measurement, 12, 231-237.

Kennedy, W. J., Jr., & Gentle, J. E. (1980). Statistical computing. New York: Marcel Dekker.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. Educational and Psychological Measurement, 56, 746-759.

Knapp, T. R., & Campbell-Heider, N. (1989). Numbers of observations and variables in multivariate analyses. Western Journal of Nursing Research, 11, 634-641.

Knuth, D. E. (1981). The art of computer programming: Vol. 2. Seminumerical algorithms (2nd ed.). Reading, MA: Addison-Wesley.

Kraemer, H. C., & Thiemann, S. (1987). How many subjects? Statistical power analysis in research. Newbury Park, CA: Sage.

Kromrey, J. D., & Hines, C. V. (1995). Use of empirical estimates of shrinkage in multiple regression: A caution. Educational and Psychological Measurement, 55, 901-925.

L'Ecuyer, P. (1988). Efficient and portable combined random number generators. Communications of the ACM, 31, 742-749, 774.

Levin, J. R. (1993). Statistical significance testing from three perspectives. Journal of Experimental Education, 61, 378-382.

Light, R. J., Singer, J. D., & Willett, J. B. (1990). By design: Planning research on higher education. Cambridge, MA: Harvard University.

Lord, F. M. (1950). Efficiency of prediction when a regression equation from one sample is used in a new sample (Research Bulletin No. 50-40). Princeton, NJ: Educational Testing Service.

Marquardt, D. W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. Technometrics, 12, 591-611.

Mason, R. L., Gunst, R. F., & Webster, J. T. (1975). Regression analysis and problems of multicollinearity. Communications in Statistics, 4, 277-292.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. Psychological Bulletin, 105, 156-166.

Miller, D. E., & Kunce, J. T. (1973). Prediction and statistical overkill revisited. Measurement

and evaluation in guidance, 6, 157-163.

Milton, S. (1986). A sample size formula for multiple regression studies. Public Opinion Quarterly, 50, 112-118.

Montgomery, D. C., & Peck, E. A. (1992). Introduction to linear regression analysis (2nd ed.). New York: John Wiley & Sons.

Mooney, C. Z. (1997). Monte Carlo simulation (Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-116). Thousand Oaks, CA: Sage.

Morgan, B. J. T. (1984). Elements of simulation. New York: Chapman and Hall.

Morris, J. D. (1981). Updating the criterion for regression predictor variable selection. Educational and Psychological Measurement, 41, 777-780.

Mosteller, F., & Tukey, J. W. (1969). Data analysis, including statistics. In G. Lindzey & E. Aronson (Eds.), The handbook of social psychology. Volume Two: Research methods (2nd ed., pp. 80-203). Reading, MA: Addison-Wesley.

Myers, R. H. (1990). Classical and modern regression with applications (2nd ed.). Boston: PWS-Kent.

Nash, J. C. (1990). Compact numerical methods for computers: Linear algebra and function minimisation (2nd ed.). New York: Adam Hilger.

Neter, J., Wasserman, W., & Kutner, M. H. (1990). Applied linear statistical models: Regression, analysis of variance, and experimental designs (3rd ed.). Homewood, IL: Irwin.

Nicholson, G. E. (1960). Prediction in future samples. In I. Olkin et al. (Eds.), Contributions to probability and statistics: Essays in honor of Harold Hotelling (pp. 322-330). Palo Alto, CA: Stanford University.

Nijsse, M. (1990). An evaluation of two techniques for constructing confidence intervals for the squared multiple correlation coefficient. Psychological Reports, 67, 1107-1116.

Norusis, M. J., & SPSS Inc. (1993). SPSS® for Windows™: Base system user's guide, release 6.0. Chicago: SPSS.

Olejnik, S. F. (1984). Planning educational research: Determining the necessary sample size. Journal of Experimental Education, 53, 40-48.

Park, C. N., & Dudycha, A. L. (1974). A cross-validation approach to sample size determination for regression models. Journal of the American Statistical Association, 69, 214-218.

Park, S. K., & Miller, K. W. (1988). Random number generators: Good ones are hard to find. Communications of the ACM, 31, 1192-1201.

Pedhazur, E. J., & Schmelkin, L. P. (1991). Measurement, design, and analysis: An integrated approach. Hillsdale, NJ: Lawrence Erlbaum Associates.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1989). Numerical recipes in Pascal: The art of scientific computing. New York: Cambridge University.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). Numerical recipes in FORTRAN: The art of scientific computing (2nd ed.). New York: Cambridge University.

Ray, A. A. (1982). SAS user's guide: Statistics, 1982 edition. Cary, NC: SAS Institute.

Ripley, B. D. (1987). Stochastic simulation. New York: John Wiley & Sons.

Robey, R. R., & Barcikowski, R. S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. British Journal of Mathematical and Statistical Psychology, 45, 283-288.

Rockwell, R. C. (1975). Assessment of multicollinearity: The Haitovsky test of the determinant. Sociological Methods and Research, 3, 308-320.

Rozeboom, W. W. (1978). Estimation of cross-validated multiple correlations: A clarification. Psychological Bulletin, 85, 1348-1351.

Rozeboom, W. W. (1981). The cross-validational accuracy of sample regressions. Journal of

Educational Statistics, 6, 179-198.

Rubinstein, R. Y. (1981). Simulation and the Monte Carlo method. New York: John Wiley & Sons.

Sampson, A. R. (1974). A tale of two regressions. Journal of the American Statistical Association, 69, 682-689.

Sawyer, R. (1982). Sample size and the accuracy of predictions made from multiple regression equations. Journal of Educational Statistics, 7, 91-104.

Schafer, W. D. (1993). Interpreting statistical significance and nonsignificance. Journal of Experimental Education, 61, 383-387.

Schmitt, N., Coyle, B. W., & Rauschenberger, J. (1977). A Monte Carlo evaluation of three formula estimates of cross-validated multiple correlation. Psychological Bulletin, 84, 751-758.

Shaver, J. P. (1993). What statistical significance testing is, and what it is not. Journal of Experimental Education, 61, 293-316.

Silvey, S. D. (1969). Multicollinearity and imprecise estimation. Journal of the Royal Statistical Society Series B, 31, 539-552.

Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size parameters. Journal of Experimental Education, 61, 334-349.

Spath, H. (1992). Mathematical algorithms for linear regression. Boston: Academic.

Stein, C. (1960). Multiple regression. In I. Olkin et al. (Eds.), Contributions to probability and statistics: Essays in honor of Harold Hotelling (pp.425-443). Palo Alto, CA: Stanford University.

Stevens, J. (1986). Applied multivariate statistics for the social sciences. Hillsdale, NJ: Lawrence Erlbaum Associates.

Stevens, J. (1996). Applied multivariate statistics for the social sciences (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Streiner, D. L. (1990). Sample size and power in psychiatric research. Canadian Journal of Psychiatry, 35, 616-620.

Tabachnick, B. G., & Fidell, L. S. (1989). Using multivariate statistics (2nd ed.). New York: HarperCollins.

Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. Journal of Experimental Education, 61, 361-377.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. Educational Researcher, 25(2), 26-30.

Uhl, N., & Eisenberg, T. (1970). Predicting shrinkage in the multiple correlation coefficient. Educational and Psychological Measurement, 30, 487-489.

Wampold, B. E., & Freund, R. D. (1987). Use of multiple regression in counseling psychology research: A flexible data-analytic strategy. Journal of Counseling Psychology, 34, 372-382.

Webster, J. T., Gunst, R. F., & Mason, R. L. (1974). Latent root regression analysis. Technometrics, 16, 513-522.

Weisberg, S. (1985). Applied linear regression (2nd ed.). New York: John Wiley & Sons.

Wherry, R. J., Sr. (1975). Underprediction from overfitting: 45 years of shrinkage. Personnel Psychology, 28, 1-18.

Willan, A. R., & Watts, D. G. (1978). Meaningful multicollinearity measures. Technometrics, 20, 407-412.

Woolf, H. B. (Ed.). (1975). Webster's new collegiate dictionary. Springfield, MA: G. & C. Merriam.

Appendix A

Derivation of the PEAR Method for Sample Size Selection

Start with the Lord formula, as presented by Uhl & Eisenberg (1970):

$$R_c^2 = 1 - \frac{N+p+1}{N-p-1}(1-R^2)$$

Multiplying both sides by (N-p-1) yields:

$$(N-p-1)(R_c^2) = (N-p-1) - (N+p+1)(1-R^2)$$

Expanding the quantities gives:

$$NR_C^2 - pR_C^2 - R_C^2 = N-p-1-N-p-1+NR^2+pR^2+R^2$$

and grouping and subtracting gives:

$$NR_C^2 - NR^2 = pR_C^2 + R_C^2 - p - 1 - p - 1 + pR^2 + R^2$$

By factoring the terms:

$$N(R_C^2 - R^2) = p(R_C^2 - 2 + R^2) + 1(R_C^2 - 2 + R^2)$$

And therefore

$$N(R_C^2 - R^2) = (p+1)(R_C^2 - 2 + R^2)$$

Multiplying both sides by (-1) and then dividing both sides by $(R^2 - R_C^2)$ gives:

$$N = (p+1)\frac{(2-R^2-R_C^2)}{(R^2-R_C^2)}$$

Let $\epsilon = R^2 - R_C^2$ and therefore $R_C^2 = R^2 - \epsilon$:

$$N = (p+1)\frac{(2-R^2-(R^2-\epsilon))}{\epsilon}$$

Finally,

$$N = (p+1)\frac{(2-2R^2+\epsilon)}{\epsilon}$$

Appendix B

Correlation Matrices for Three and Seven Predictors

Table D.1

Correlation Matrices for Three Predictors

| Matrix Condition | $\rho^2$ | | $y$ | $x_1$ | $x_2$ |
|---|---|---|---|---|---|
| Orthogonal | .40 | $x_1$ | .292 | | |
| | | $x_2$ | .270 | .000 | |
| | | $x_3$ | .492 | .000 | .000 |
| | .25 | $x_1$ | .257 | | |
| | | $x_2$ | .257 | .000 | |
| | | $x_3$ | .343 | .000 | .000 |
| | .10 | $x_1$ | .088 | | |
| | | $x_2$ | .137 | .000 | |
| | | $x_3$ | .271 | .000 | .000 |
| Non-Multicollinear | .40 | $x_1$ | .292 | | |
| | | $x_2$ | .270 | .265 | |
| | | $x_3$ | .492 | .080 | -.192 |
| | .25 | $x_1$ | .257 | | |
| | | $x_2$ | .257 | -.206 | |
| | | $x_3$ | .343 | .800 | -.277 |
| | .10 | $x_1$ | .088 | | |
| | | $x_2$ | .137 | .610 | |
| | | $x_3$ | .271 | .376 | .098 |
| Moderately Multicollinear | .40 | $x_1$ | .292 | | |
| | | $x_2$* | .270 | .809 | |
| | | $x_3$ | .492 | .256 | .614 |
| | .25 | $x_1$ | .257 | | |
| | | $x_2$* | .257 | .709 | |
| | | $x_3$ | .343 | .131 | .683 |
| | .10 | $x_1$ | .088 | | |
| | | $x_2$* | .137 | .812 | |
| | | $x_3$ | .271 | .316 | .704 |
| Extensively Multicollinear | .40 | $x_1$ | .292 | | |
| | | $x_2$* | .270 | .240 | |
| | | $x_3$* | .492 | .621 | .846 |
| | .25 | $x_1$ | .257 | | |
| | | $x_2$* | .257 | .680 | |
| | | $x_3$* | .343 | .741 | .976 |
| | .10 | $x_1$* | .088 | | |
| | | $x_2$* | .137 | .907 | |
| | | $x_3$ | .271 | .624 | .595 |

* indicates predictor with $VIF > 5.0$ (i.e., involved in multicollinearity)

## Appendix C

## Stem-and-Leaf Plots of the Precision Efficacy Accuracy of Several Sample Size Methods

These plots were adapted from Brooks and Barcikowski (1995). The accuracy criterion used in that study for these results was $.75 \leq PE \leq .85$. Those leaves which represent accurate results have been boldfaced and underlined. For every plot, the stem width is 0.1000. Each leaf represents one case.

PEAR Method (Precision Power by Brooks & Barcikowski, 1995)

| Frequency | Stem & | Leaf |
|---|---|---|
| .00 | 0 . | |
| .00 | 1 . | |
| .00 | 2 . | |
| .00 | 3 . | |
| .00 | 4 . | |
| .00 | 5 . | |
| .00 | 6 . | |
| 1.00 | 7 . | **9** |
| 19.00 | 8 . | **0011111111112222222** |
| .00 | 9 . | |

Sawyer (1982)

| Frequency | Stem & | Leaf |
|---|---|---|
| .00 | 0 . | |
| .00 | 1 . | |
| .00 | 2 . | |
| 1.00 | 3 . | 9 |
| 6.00 | 4 . | 022368 |
| 3.00 | 5 . | 011 |
| 5.00 | 6 . | 23455 |
| 1.00 | 7 . | **9** |
| 4.00 | 8 . | **0112** |
| .00 | 9 . | |

Predictive Power Method (Brooks & Barcikowski, 1994)

| Frequency | Stem & | Leaf |
|---|---|---|
| .00 | 0 . | |
| .00 | 1 . | |
| .00 | 2 . | |
| .00 | 3 . | |
| .00 | 4 . | |
| .00 | 5 . | |
| 2.00 | 6 . | 69 |
| 13.00 | 7 . | 01134**55667999** |
| 5.00 | 8 . | **00001** |
| .00 | 9 . | |

*30:1* subject-to-predictor ratio (Pedhazur & Schmelkin, 1991)

| Frequency | Stem & | Leaf |
|---|---|---|
| .00 | 0 . | |
| .00 | 1 . | |
| .00 | 2 . | |
| 3.00 | 3 . | 166 |
| 2.00 | 4 . | 37 |
| .00 | 5 . | |
| 1.00 | 6 . | 9 |
| 3.00 | 7 . | 4**69** |
| 1.00 | 8 . | **0** |
| 10.00 | 9 . | 0122277777 |

Park and Dudycha (1974)

| Frequency | Stem & | Leaf |
|---|---|---|
| .00 | 0 . | |
| .00 | 1 . | |
| .00 | 2 . | |
| .00 | 3 . | |
| 1.00 | 4 . | 4 |
| 5.00 | 5 . | 11349 |
| 2.00 | 6 . | 69 |
| 3.00 | 7 . | 01**9** |
| 6.00 | 8 . | **455**679 |
| 3.00 | 9 . | 000 |

*50 + 8p* conventional rule (Green, 1991)

| Frequency | Stem & | Leaf |
|---|---|---|
| .00 | 0 . | |
| 2.00 | 1 . | 06 |
| 2.00 | 2 . | 39 |
| 1.00 | 3 . | 4 |
| .00 | 4 . | |
| 2.00 | 5 . | 18 |
| 2.00 | 6 . | 58 |
| 1.00 | 7 . | 2 |
| 4.00 | 8 . | **14**89 |
| 6.00 | 9 . | 134667 |

BEST COPY AVAILABLE

15:1 N:p ratio (Stevens, 1996)

| Frequency | Stem & | Leaf |
|---|---|---|
| .00 | 0 . | |
| 5.00 | 1 . | 56777 |
| .00 | 2 . | |
| .00 | 3 . | |
| 2.00 | 4 . | 39 |
| 2.00 | 5 . | 39 |
| 1.00 | 6 . | 1 |
| 1.00 | 7 . | **9** |
| 4.00 | 8 . | **1355** |
| 5.00 | 9 . | 34445 |

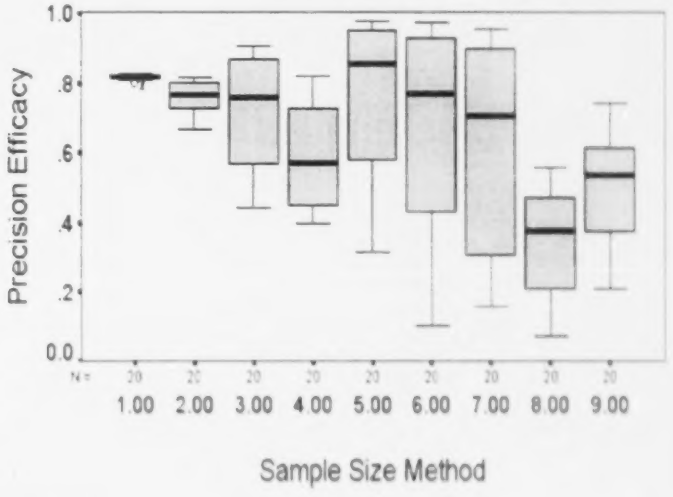
Gatsonis and Sampson (1989)

| Frequency | Stem & | Leaf |
|---|---|---|
| .00 | 0 . | |
| .00 | 1 . | |
| 3.00 | 2 . | 125 |
| 3.00 | 3 . | 667 |
| 3.00 | 4 . | 299 |
| 5.00 | 5 . | 34668 |
| 4.00 | 6 . | 1156 |
| 2.00 | 7 . | 04 |
| .00 | 8 . | |
| .00 | 9 . | |

Cohen (1988)

| Frequency | Stem & | Leaf |
|---|---|---|
| 1.00 | 0 . | 7 |
| 3.00 | 1 . | 169 |
| 2.00 | 2 . | 01 |
| 5.00 | 3 . | 02468 |
| 6.00 | 4 . | 034588 |
| 3.00 | 5 . | 025 |
| .00 | 6 . | |
| .00 | 7 . | |
| .00 | 8 . | |
| .00 | 9 . | |

Comparison boxplots of the levels of precision efficacy for the methods for the 20 conditions tested by Brooks and Barcikowski (1995). Method 1.00 is the PEAR Method; 2.00 is the Predictive Power method; 3.00 is the Park and Dudycha (1974) method; 4.00 is the Sawyer (1982) method; 5.00 is the *30:1* ratio; 6.00 is the $50 + 8p$ method; 7.00 is the *15:1* ratio; 8.00 is Cohen's (1988) method; and 9.00 is Gatsonis and Sampson's (1989) method.

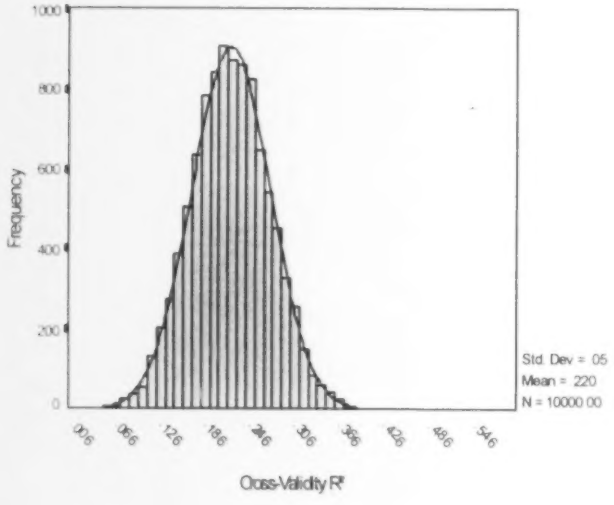



BEST COPY AVAILABLE

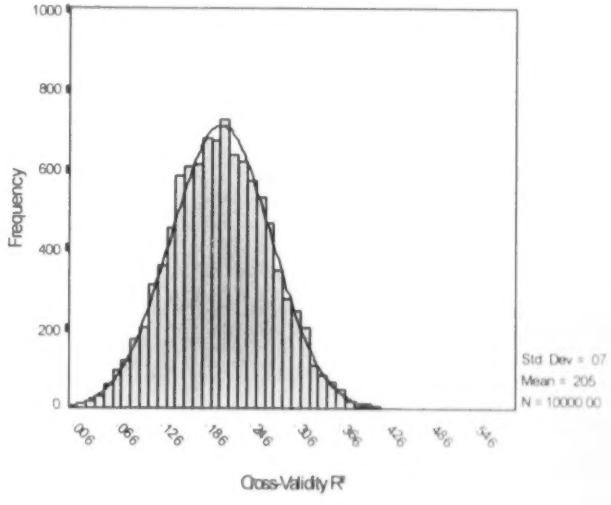
## Appendix D
## Histograms of Cross-Validity $R^2$ for Seven Predictors at Effect Size $\rho^2 = .25$

These figures were created from data collected for each of the 10,000 samples at effect size $\rho^2 = .25$ with seven predictors in the orthogonal multicollinearity condition. A curve that represents the normal distribution is superimposed on the cross-validity $R^2$ distribution for each of the following graphs.

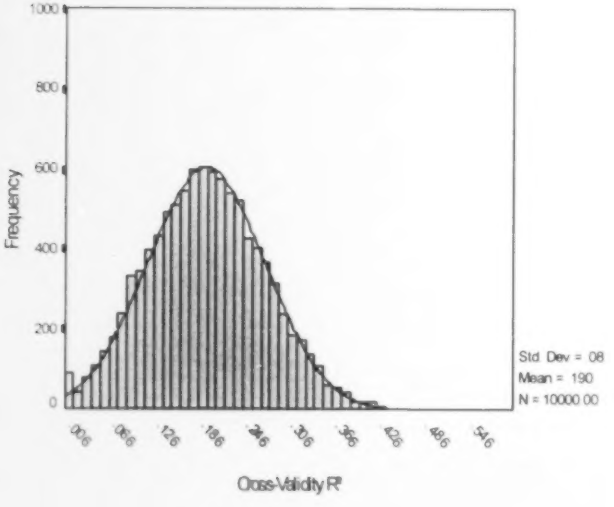
a priori $PE = .80$, 7 predictors




a priori $PE = .70$, 7 predictors




a priori $PE = .60$, 7 predictors




15:1 subject-to-predictor ratio, 7 predictors